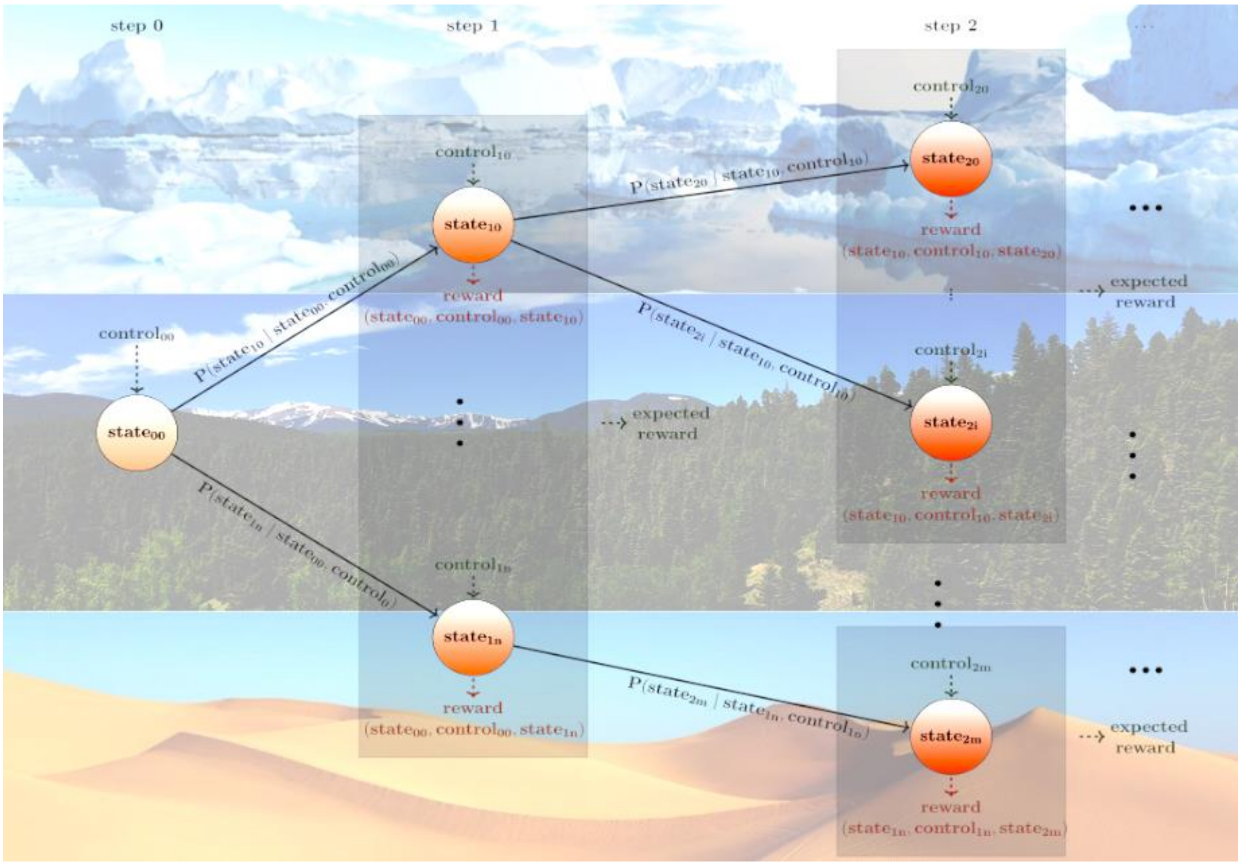
TIPPING POINTS IN THE EARTH SYSTEM

WP6 Domain-specific-language with basic notions about Tipping Points: D6.2 (D28)


step 0
step 1
step 2
...
control₀₀
state₀₀
P(state₁₀ | state₀₀, control₀₀)
P(state₁ₙ | state₀₀, control₀)
control₁₀
state₁₀
reward
(state₀₀, control₀₀, state₁₀)
control₁ₙ
state₁ₙ
reward
(state₀₀, control₀₀, state₁ₙ)
expected
reward
P(state₂₀ | state₁₀, control₁₀)
P(state₂ᵢ | state₁₀, control₁₀)
P(state₂ₘ | state₁ₙ, control₁ₙ)
control₂₀
state₂₀
reward
(state₁₀, control₁₀, state₂₀)
control₂ᵢ
state₂ᵢ
reward
(state₁₀, control₁₀, state₂ᵢ)
control₂ₘ
state₂ₘ
reward
(state₁ₙ, control₁ₙ, state₂ₘ)
expected
reward
expected
reward

TiPES: Tipping Points in the Earth System is *a Research and Innovation action (RIA) funded by the Horizon 2020 Work programme topics "Addressing knowledge gaps in climate science, in support of IPCC reports" Start date: 1st* September 2019. End date: 31st August 2023.

The TiPES project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 820970.

About this document

Deliverable: D6.2
Work package in charge: WP6 Uncertainty and accountable policies
Actual delivery date for this deliverable: Project-month 36
Dissemination level:
The general public (PU)

Lead author(s)
Potsdam Institute for Climate Impact Research (PIK): Nicola Botta, Nuria Brede

Other contributing author(s)
Université catholique de Louvain (UCL): Michel Crucifix, Marina Martínez Montero

Reviewer(s)
University of Copenhagen (UCPH): Peter Ditlevsen

Visit us on: www.tipes.dk

Follow us on Twitter: @TiPES_H2020

Access our open access documents in Zenodo:
https://zenodo.org/communities/tipes/

Disclaimer: This material reflects only the author’s view, and the Commission is not responsible for any use that may be made of the information it contains.

Index

Summary for publication

In climate policy, decisions have to be taken *sequentially*, under *uncertainty* and based on imperfect data. Policy makers can control greenhouse gas emissions only indirectly through carbon taxes, emission rights trading schemes, etc. Credible policy advice therefore has to account for uncertainties, both about our understanding of how the climate system responds to anthropogenic forcing but also, and more importantly, about the capability of policy makers to effectively implement emissions reduction plans.

Viable climate policies need to balance conflicting interests and be robust under deep uncertainty. Rationalising climate policies requires, among others, understanding and quantifying which climate decisions matter most, how uncertainties affect optimal decisions and how current decisions may shrink (or widen) the decision space of future generations.

This deliverable reports on an ontology of climate science notions for tipping point research and on a *domain-specific language* (*DSL*)[1] for policy advice under deep threshold uncertainty. The ontology and the domain-specific language have been designed to assist the specification and the verified solution of stylised climate decision problems. Applications of the ontology and of the DSL are discussed in related publications.

In these applications, the DSL serves two main purposes. First, it clarifies and assigns meanings to notions which are used ambiguously in tipping point research. Second, it provides useful abstractions for specifying and solving decision problems under uncertainty. This is crucial for delivering *accountable* policies, that is, policies that are verified against their specification.

The first purpose also motivates the development of the ontology. We have formalised and annotated the most compelling notions of *climate sensitivity*, *commitment*, *abrupt change*, *tipping point*, *tipping element* and *early warning signal* available in the literature. For some of these notions, we have discovered computational patterns similar to those presented in [Ion09] for the notion of *vulnerability*.

With respect to the second purpose, we have developed an abstraction layer via DSL extensions of the framework for specifying and solving *monadic sequential decision problems* (*SDPs*) of Botta et al. [BIJ17]. The notion of *monadic SDP* can be understood as formalising a generic notion of sequential dynamics under uncertainty, which can be instantiated to recover concrete instances. Among others, the familiar ones: deterministic, non-deterministic, and stochastic dynamical systems, which all can be used to study and quantify the consequences of uncertainty.

Specifically, we have extended the framework with generic measures of responsibility and with a DSL for transparently expressing goals of decision making. This allows to specify climate decision problems in terms of value judgments following ideas of planetary boundaries and safe operational spaces [Roc+09, Hei+16]. We have applied this DSL to obtain accountable measures of how much climate decisions under uncertainty matter. In this application we have also developed a method for encoding transition functions of stochastic decision problems in a modular way from Bayesian belief networks.

A major contribution to optimal decision theory has emerged from the need to apply the Botta et al. framework to non-standard instances of monadic sequential decision problems studied within WP6. In this contribution, we have formulated sufficient and mutually independent conditions on

[1] Different from a general-purpose computer language, a domain-specific language (DSL) is specialized to a particular domain of application. In our case, the domain of application is that of climate science, focussing on notions that arise in the context of tipping point research and climate policy.

combinations of measures, monads and value aggregation functions for the framework's generic backwards induction to be correct. The proof of the result is implemented in the Idris computer language and machine checked.

The main results achieved in this deliverable are

- an ontological review and operational description of key notions of tipping point research (research report [BBCM22a]),
- a DSL for monadic decision problems, responsibility under uncertainty and tipping point notions (research report [BBCM22b]),
- a new correctness result for generic backward induction (publications [BB21, Bot+21a])
- a novel method for estimating how much climate decisions under uncertainty matter (submitted paper [Bot+21b])

Work carried out

Details of work carried out

The details of the work carried out for this deliverable have been assembled in two stand-alone reports that are attached as Appendix A and B to this document and are publicly available on Zenodo (see dissemination section [BBCM22a] and [BBCM22b]). Here we just outline the contents of these two reports.[2]

Climate sensitivity, commitment and abrupt change: toward an ontology for climate TP research

In the report [BBCM22a] (Appendix A) we review relevant notions in the context of tipping point (*TP*) research and sketch how to organise them in a prototype ontology. This is the first part of work task T6.1.1 and objective O6.1. Particular attention is given to the following notions:

- different variants of *climate sensitivity* [KR17] (Appendix A, Section 4),
- *(climate change) commitment* [Wig05] (Appendix A, Section 5),
- *abrupt change* [All+03], *tipping point* and *tipping element* [Len+09] (Appendix A, Section 6),
- *early warning signal* [Sch+09] (Appendix A, Section 7)

We tackle these notions from the point of view of dynamical systems theory as originally proposed in [Ion09][3] and formalise climate sensitivity and commitment in terms of generic schemes that can be instantiated with the variants found in the literature. This shows that climate sensitivity and commitment are in fact closely related notions.

For abrupt change, tipping point, tipping element and early warning signal, we review and annotate compelling definitions like for example that of [Len+09] (Appendix A, Section 6). For each notion, we present a list of selected papers and a bibliography. This allows to trace its respective historical development and the most relevant contributions.

[2] Concerning the dissemination of the results presented in the two reports see also below in “Main results achieved” and “How we are going to ensure the uptake of the deliverable by the targeted audience”.

[3] In this study, Ionescu uses functional programming inspired formal methods to clarify the notion of *vulnerability* and proposes a generic operational description that can be instantiated to different variants of the notion found in the literature.

At the end, we sketch a framework for organising the different notions and their relationships in a formal ontology. In this framework, important concepts like time series of observational data or dynamical systems and their trajectories arise as instances of generic notions of *sequential data* and *sequential data producers*.

A DSL for Monadic Decision Problems, Responsibility under Uncertainty and Tipping Point Notions

In the report [BBCM22b] (Appendix B) we describe the modifications and extensions that we have implemented as domain-specific language extensions in the IdrisLibs framework of [BIJ17]. This is the second part of work task T6.1.1 and objective O6.1. Based on the central notions of *monadic decision process* and *monadic sequential decision problem (MSDP),* in which the category-theoretical structure of a *monad* [Mac78, Wad92] captures a generic notion of uncertainty [Ion09, BIJ17] (for example non-deterministic or stochastic uncertainty [Gir81, EK06]), we have decided to work with a lightweight version of the [BIJ17] theory in a trade-off between expressivity and user-friendliness.

As a first step, we discuss what it means for solutions of MSDPs to be optimal and show under which conditions we can prove that the generic backward induction algorithm of the framework indeed computes optimal solutions (Appendix B, Section 4). With this we also address the question posed in the description of TiPES cross cutting Theme 4 of what it means for decisions to be optimal under unavoidable political uncertainty and imperfect information. This part of the work has led to the publications [Bot+21a] and [BB21].

We furthermore describe the development of generic responsibility measures and a syntax for transparently describing the goals of decision making (Appendix B, Section 7, addressing work task T6.1.3 and objective O6.4). Their usage is illustrated with a stylised stochastic emission problem (Appendix B, Section 5). We also show how to modularly describe the transition function of the underlying decision process with conditional probabilities in the sense of Bayesian belief networks. This part of the report is based on the submitted paper [Bot+21b] but contains some simplifications and extensions.

Given that the application of formal methods to climate science and to climate policy advice is far from mainstream approaches (which include integrated assessment modelling or standard scenario simulation), we have assembled some thoughts about “climate science and verified programming”, “climate science and climate policy” and “decision theory and climate policy” in three short notes [BBCM21a-c] which can be found in Appendix C-E.

The main responsibility for the work carried out in the context of this deliverable was located at PIK. The UCL partner has provided advice and feedback concerning domain-specific notions, especially in the context of the ontological literature review. A novel notion of *lost options commitment* and the underlying climate model have been conceived by UCL in the context of deliverable D6.3 and the Idris formalisation has been carried out by PIK.

Deviations from the DoA, difficulties in the implementation

Already at the kick-off meeting in Paris in September 2019 it became clear that it would be difficult to adopt a consensual, clear-cut definition of the notion of “tipping point”. This situation did not perceivably change in the following. In general, assembling a prototype ontology of tipping point notions for T6.1.1 turned out to be much more time-consuming and challenging than expected. This was due to several different factors. First, we had to acknowledge that there is no consensual definition of tipping point in the literature. “Semantic confusion” around the notion of tipping point is widely acknowledged in the literature and, according to some [RN09, Rus15], the tipping point metaphor can act as a rhetorical device, despite ambiguity of its technical meaning. Even within the TiPES consortium, an unambiguous definition of tipping point appeared to escape consensus.

This situation was moreover complicated by the diverse backgrounds of the work package members, with the main responsibility for this deliverable located with the non-climate scientist members of WP6. Due to the COVID-19 pandemic, more extended "on-site" personal interactions between the work package members within WP6, but also with other work packages were made impossible.

Despite these difficulties, the work carried out for this deliverable remains firmly anchored in the science surrounding tipping points, and consistent with work carried out in other work packages. Namely, the notions reviewed and formalised here will be used in deliverable D6.3, where problems are posed and solved with a climate model developed within the framework of WP6 (UCL partner). This model is called "SURFER" [Mar+22], it instantiates the notion of tipping point in ice dynamics and facilitates implementing the notions of "desirable" and "undesirable states".

Main results achieved

The main results achieved with this deliverable consist, on the one hand, in an ontological review and operational description of certain key notions of tipping point research; on the other hand, in an extension of the Botta et al. IdrisLibs framework with several new domain-specific language elements for studying climate policy problems, in particular generic measures of responsibility, a syntax for expressing the goals of decision making transparently and operational definitions for studying commitment. They are essential first steps in providing an abstraction layer to narrow the gap between problem specification and implementation, and to allow for the use of formal verification to improve accountability.

In the publications [Bot+21a] and [BB21] we report on the theoretical and technical foundations of the notion of optimality and method of optimization implemented in the framework.
We consider that programs used for policy advice must be proved to provide correct solutions once the problem they are meant to solve is specified. This is a necessary condition for making the policy accountable. We thus derive sufficient conditions that allow us to ensure this kind of correctness for generic (monadic) backward induction as a solution method for monadic sequential decision problems. This is particularly relevant to ensure correctness when studying non-standard instances of sequential decision problems using measures of uncertainty not commonly used in control theory, e.g., motivated by the paradigm shift from cost-benefit to risk-opportunity analysis discussed in [Sha+21].

In the submitted paper [Bot+21b], we develop generic measures of responsibility (for sequential decision processes under uncertainty) and a small syntax for expressing the goals of decision making in a transparent and fair way. The proposed measures of responsibility are consistent with three conditions under which "a person can be ascribed responsibility for a given outcome" which have been put forward in the literature [BvH18]: *avoidance*, *agency,* and *causal relevance*. As a first application, this theory of responsibility under uncertainty has been applied to study the temporal evolution of responsibility in a highly stylised GHG emission decision problem.

Not all results have already been disseminated in peer-reviewed publications or submitted manuscripts. This concerns primarily the content of the report [BBCM22a] (Appendix A) which

provides preliminary results for an ontology of operationalised key notions in the context of tipping point research which we plan to submit as a contribution to EarthArxiv (open source).
Other (though minor) results that have not yet been published are reported in [BBCM22b] (Appendix B). These are extensions of the IdrisLibs framework that prepare the theoretical underpinnings for upcoming work contributing to TiPES D6.3 (see [BBCM22b, Section 3]) and for future work building on the ideas presented in [Bot+21b] (see [BBCM22b, Sections 5-7]).

Progress beyond the state of the art

While climate models can, up to a certain extent, be validated on the basis of indirect observations of past climates (palaeoclimatology) and of a growing amount of direct observations, and the (conditional) probabilities of different climate change scenarios (for given anthropogenic forcings) can be estimated, there is no consensual approach for assessing the effects of climate change on societies and for reliably estimating the feedback of climate change on anthropogenic forcing.

Because of this asymmetry, climate science has been so far incapable of providing advice on matters of climate policy that is *accountable*: decision makers do not precisely know what kind of guarantees they can expect from implementing the advice received.

State of the art *Integrated assessment models (IAMs)* of climate change of the kind discussed in [Nor18] have been widely applied to inform decision making but they have also been criticised, mainly because of three reasons: 1) their lack of predictive capability; 2) their reliance on cost-benefit analysis and marginality assumptions and 3) their focus on deterministic sequential decision problems.

Another important aspect which has not been satisfactorily addressed by state of the art approaches is that of *responsibility,* and one of the objectives of TiPES WP6 was indeed to "Develop and apply influence and responsibility measures for accountable decision making (O4)".

While notions of *ex-post responsibility* are crucial for the attribution of liabilities, e.g., for past GHG emissions, planning in matters of climate policy needs to be informed by *ex-ante measures of how much decisions matter*.

We know that climate decisions which are taken (or delayed) now and in the next decades, e.g., on greenhouse gas (GHG) emissions, will be crucial for the upcoming generations. But do current decisions matter more or less than decisions to be taken in, say, one or two decades? Can an agent be held responsible for (performing or for failing to perform) actions that matter very little? And what does it precisely mean “to matter”, for decisions that are taken under deep uncertainty and imperfect information?

The work reported in this deliverable goes beyond the state of the art in the following ways:

- The ontology and the DSL are directed towards applying formal methods to the specification and solution[4] of decision problems under deep uncertainty. We consider that this is an essential contribution. It can facilitate and clarify the dialog with policymakers, and we argue it could substantially enhance the legal foundations related to the climate agreements. Indeed, the work is a step towards reaching a new level of *accountability:* thanks to the underlying foundation of Dependent Type Theory, the results computed within our framework are machine-checked to be logical consequences of the assumptions made. This

[4] notably, by employing a computer-verified algorithm

allows for a much higher degree of confidence in the correctness of results than can be obtained with conventional programming languages.

- Being built around a generic notion of sequential decision problems under uncertainty, our theory is much more flexible and re-usable for different combinations of uncertainty, measures, and utility functions than the common state of the art approaches to cost-benefit analysis.
- Backward induction [Bel1957] is the go-to efficient method for solving sequential decision problems. With the formulation of correctness conditions for the generalised monadic backward induction we provide reasonably simple criteria for when this method can be used for efficiently solving sequential decision problems involving non-standard combinations of uncertainties, measures and value structures (which is crucial when moving from classical cost-benefit to a more informative risk-opportunity analysis paradigm has been recently advocated e.g., by [Sha+21]). To our knowledge these criteria had not been formulated in this generality before and our correctness theorem is a genuinely new theoretical result, of which we in addition provide a formalised and machine-checked proof. It should be noted that not only the optimization algorithm of backward induction is interesting as efficient algorithm, but also its underlying *value function* which provides an efficient method for assessing the utility of a given policy sequence or scenario. The conditions formulated in our result for the applicability of monadic backward induction also apply for the applicability of the efficient value function.
- We propose a novel method for estimating how much decisions matter under monadic uncertainty. This method is generic and suitable for measuring responsibility in finite horizon sequential decision problems with monadic uncertainty. It fulfils *fairness* requirements and three natural conditions for responsibility measures (*agency*, *avoidance* and *causal relevance)* that have been formulated in the literature as requirements on responsibility measures.

Besides the technical challenges, this research programme is also ambitious because the methods and mathematical approaches followed here are not standard in climate science. This generates a context prone to communication challenges, misunderstanding, and perhaps controversies about methods. Specifically, we have followed the rationale that taking good decisions in presence of uncertainty requires transparency of assumptions and simple, understandable models. Yet, the reality is complex, and climate scientists are inclined to use large simulators which defy human comprehension. We have maintained that in this context, simple models, with transparent and simple assumptions (but calibrated on more complex models) are better suited for accountable decision making, because the related assumptions can more easily be enumerated and agreed upon on the basis of expert guidance. Our work has been entirely guided by these principles.

Impact

The work reported in this deliverable contributes to the expected impacts of TiPES of **"providing added-value to decision and policy makers"** with

- a framework for decision making which allows to systematically account for imperfect information, for uncertainty about a decision's outcome and even its effective implementation in the first place. Such uncertainties are unavoidable in decision making in the context of climate science, e.g., coming from uncertainty about the climate system itself, economical and societal issues;
- a language for climate scientists and policy experts that allows making the assumptions behind

policy advice transparent, and
- a methodology for decision making and attribution of responsibility under uncertainty and imperfect information.

Although the current impact of this work, if measured today by audience and citations, may be limited, we consider that it poses the foundations for a paradigm shift. To show this, we draw the reader's attention to current difficulties and even controversies related to the validity of mainstream approaches for assisting decision making. Current cost-benefit approaches based on integrated assessment models -- widely based on a neo-classical economic paradigm -- have important problems:

- they fail to address the long-term legacy of current decisions (the reward of future states is typically discounted);
- models are complex, thus hard to comprehend (integrated assessment models contain many parameters and assumptions), and
- defining "optimal decisions" depends on arbitrarily weighing distinct and potentially conflicting value judgements about different aspects of climate change and climate intervention. In formal terms, fixing prices is problematic.

Hence, we pose the diagnostic that minimising the cost predicted by integrated assessment models is unlikely to guide the policymaker towards decisions that are accountable and well understood. This diagnostic was shared and stressed by most participants of the virtual workshop *"Challenges and new directions in risk analysis, decision making and policy advice for climate change"* organised by WP6 (see below). The audience success of this workshop is a sign of the importance and potential impact of the work carried out under WP6.

The ontology, DSL and monadic framework provide healthy bases for tackling the difficulties enumerated above. This will become evident in the work carried out under D6.3 which will, for example, implement and use the notion of commitment.

Lessons learned and links built

Lessons learned

Optimal decision making about climate change is not an ordinary optimal decision problem, because of the long-term legacy of current decisions, the values at stake, uncertainties, semantic ambiguities, and the complex ethical context. We certainly have a better appreciation of these difficulties than we had two years ago.

We have however been comforted in our approach to use simple models with transparent assumptions. We have clarified that advising decision making requires more than a dynamical system framework. The dynamical system needs to be complemented by a set of controls, a value structure, and an explicit specification of uncertainties. On this basis, we have learned how to specify a responsibility measure and how to implement it in a computer language with computer-checked proofs of the results obtained.

Developing the ontology and the DSL have been time consuming tasks, in fact more than anticipated. But they have convinced us that it is enlightening to have at hand a consistent, well thought-out algebraic structure linking notions around decision making, climate change and tipping point. It allows us to link different definitions or problems as particular cases of a general framework. It also helped us to frame policy problems that at least partly address the difficulties emerging from the long-term legacy of decisions, the conflict of values and the complexity of models. This, we hope, will appear clearer in the following developments of the project.

Furthermore, the interdisciplinary context associating two physicists (UCL partner), a computer scientist and an engineer has generated a fruitful context and created opportunities for articles not originally planned, such as [Bot+21a] and [BB21].

Links built

- This deliverable is strongly linked to the other deliverables of WP6: The work on the correctness of monadic backward induction was inspired from discussions in the context of D6.1, and it provides a robust base of unambiguously defined notions for the work carried out towards D6.3 and D6.4.
- Resulting from a discussion during the TiPES M18 General Assembly, the note on “Decision theory and climate policy” [BBCM21c] (Appendix E) was written as potential contribution in the context of WP7’s deliverable D7.2.
- To promote the use of formal methods and foster discussion among TiPES members, we circulated a note on “Climate science and verified programming” [BBCM21a] on the TiPES mailing list.
- We organised the following workshops (both in association with this deliverable and the forthcoming deliverable D6,3) that stimulated interesting discussions:
    - A virtual TiPES crosscutting Theme 4 workshop with invited speakers Claudia Wieners (Utrecht University) and our external partner Patrik Jansson (Chalmers IoT). This workshop led to subsequent interactions with Claudia Wieners and hopefully more cooperation in the future.
    - A 1-week virtual internal WP6 workshop with attendance of our external partners Patrik Jansson and Cezar Ionescu (TH Deggendorf).
    - A thematically broader virtual workshop *“Challenges and new directions in risk analysis, decision making and policy advice for climate change”* with invited speakers Simon Sharpe (University College London, IIPP), Steve Keen (University College London, ISRS), Ted Shepherd (University of Reading) and Thomas Stocker (University of Bern, OCCR, and lead of TiPES WP7). The workshop was very well received and sparked lively side discussions not only during the discussion session, but also in the Zoom chat during all of the workshop. Hopefully, these discussions will be continued in presence at the conference on Tipping Points that will be organised in September 2022 by our TiPES partners at the University of Exeter.
- We had frequent online meetings with our external partners and co-authors Patrik Jansson and Cezar Ionescu, and with our co-authors Tim Richter (University of Potsdam) and Zheng Li (Northeastern University & Arima Inc.). These interactions resulted in the joint papers [Bot+21a] and [Bot+21b].

- Some of the ideas developed in Appendix B [BBCM22b] are currently applied in a collaboration between Nicola Botta (PIK, TiPES WP6), Patrik Jansson and Nick Smallbone (CSE, Chalmers) and the Plasma Theory group at Chalmers University of Technology on the mitigation of runaway currents in tokamak fusion devices.

Relations to the TiPES crosscutting themes

Of the themes indicated in the Description of the Action, part B, Section 1.1, this deliverable contributes to

Theme 1. Tipping Elements in data and models:

We reviewed definitions of the notions *tipping point*, *tipping element* and *abrupt change* to be found in the literature as basis for possible later formalization. This review is Section 6 of the report attached to this document as Appendix A.

Theme 2. Climate response and Early Warning Signals:

We reviewed definitions of different notions of *climate sensitivity*, *climate change commitment* and *early warning signals* to be found in the literature as basis for possible later formalization. For climate sensitivity and climate change commitment we moreover proposed generic operational descriptions. This work can be found in Sections 4, 5 and 7 of the report attached to this document as Appendix A.

Theme 4. Data and decisions:

In climate policy, decisions have to be taken sequentially, under strong political uncertainty and on the basis of imperfect data. As the recent pandemic and the more actual energy crisis make clear, this typically implies compromising between conflicting interests, often trading ideal solutions for viable ones, and trying to avoid the worst. Under these conditions it becomes crucial to be able to 1) to precisely formulate the *goals* of decision making and 2) to assess how much specific decisions *matter* for achieving these goals. The work carried out for this deliverable addresses these issues and provides dependable solutions. For example, in [Bot+21b] (reported in Appendix B, [BBCM22b]) by developing a new methodology for assessing how much decisions under uncertainty matter.

Contribution to the top-level objectives of TiPES

Of the objectives and specific goals indicated in the Description of the Action, part B, Section 1.1, this deliverable contributes to

Objective 5-Bridge the gap between climate science and policy advice

It contributes to Specific Objective 5.1. by reviewing definitions in the literature of key tipping point notions (Appendix A) and developing a domain specific language providing unambiguous meaning to the notions used for decision making under uncertainty (Appendix B).

It contributes to Specific Objective 5.2. by adding to the understanding of the impact of generic uncertainty on the mathematical problem of computing optimal policy sequences with the result reported in Appendix B, Section 4.

References (Bibliography)

[All+03] Richard B Alley et al. “Abrupt climate change”. In: science 299.5615 (2003), pp. 2005–2010.

[Bot+21c] Nicola Botta et al. IdrisLibs. https://gitlab.pik-potsdam.de/botta/IdrisLibs. 2016–2021.

[BJI17] Nicola Botta, Patrik Jansson, and Cezar Ionescu. “Contributions to a computational theory of policy advice and avoidability”. J. Funct. Program., 27:e23, 2017.

[BvH18] Matthew Braham and Martin van Hees. “Voids or Fragmentation: Moral Responsibility For Collective Outcomes”. The Economic Journal, 128(612):F95–F113, 01 2018.

[EK06] Martin Erwig and Steve Kollmansberger. “Functional Pearls: Probabilistic functional programming in Haskell”. In: J. Funct. Program. 16.1 (2006), pp. 21–34.

[Gir81] M. Giry. “A categorial approach to probability theory”. In: Categorical Aspects of Topology and Analysis. Vol. 915. Lecture Notes in Mathematics. Berlin: Springer, 1981, pp. 68–85.

[Hei+16] Jobst Heitzig et al. “Topology of sustainable management of dynamical systems with desirable states: from defining planetary boundaries to safe operating spaces in the Earth system”. In: Earth System Dynamics 7.1 (2016), pp. 21–50.

[Ion09] Cezar Ionescu. “Vulnerability Modelling and Monadic Dynamical Systems”. PhD thesis. Freie Universität Berlin, 2009. url: https://d-nb.info/1023491036/34.

[KR15] Reto Knutti and Maria AA Rugenstein. “Feedbacks, climate sensitivity and the limits of linear models”. In: Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 373.2054 (2015), p. 20150146.

[KRH17] Reto Knutti, Maria AA Rugenstein, and Gabriele C Hegerl. “Beyond equilibrium climate sensitivity”. In: Nature Geoscience 10.10 (2017), pp. 727–736.

[LS07] Timothy M Lenton et al. “Tipping elements in the Earth’s climate system”. In: Proceedings of the national Academy of Sciences 105.6 (2008), pp. 1786–1793.

[Mac78] Saunders MacLane. Categories for the Working Mathematician. 2nd. Graduate Texts in Mathematics. Springer, 1978.

[Mar+22] Marina Martínez Montero et al. “SURFER v1.0: A flexible and simple model linking emissions to sea level rise”. Submitted to Geoscientific Model Development. Preprint: https://doi.org/10.5194/egusphere-2022-135.

[Nor18] Nordhaus, William. 2018. “Evolution of Modeling of the Economics of Global Warming: Changes in the DICE Model, 1992–2017.” Climatic Change 149 (4): 623–40.

[Roc+09] Johan Rockström et al. “A safe operating space for humanity”. In: nature 461.7263 (2009), pp. 472–475.

[RN09] Chris Russill and Zoe Nyssa. “The tipping point trend in climate change communication”. In: Global environmental change 19.3 (2009), pp. 336–344.

[Rus15] Chris Russill. “Climate change tipping points: origins, precursors, and debates”.
In: Wiley Interdisciplinary Reviews: Climate Change 6.4 (2015), pp. 427–434.

[Sch+09] Marten Scheffer et al. “Early-warning signals for critical transitions”.
In: Nature 461.7260 (2009), pp. 53–59.

[Sha+21] Simon Sharpe et al. “Deciding how to decide: Risk-opportunity analysis as a generalisation of cost-benefit analysis”.
TR UCL Institute for Innovation and Public Purpose, Working Paper Series (IIPP WP 2021/03).

[Wad92] Wadler, P. “Monads for functional programming”.
NATO ASI Series 118 (1992), pp. 233–264. Springer.

[Wig05] T. M. L. Wigley. “The Climate Change Commitment”.
In: Science 307.5716 (2005), pp. 1766–1769. doi: 10.1126/science.1103934.

Dissemination and exploitation of TiPES results

Dissemination activities

Type of dissemination activity	Name of the scientist (institution), title of the presentation, event	Place and date of the event	Estimated budget	Type of Audience	Estimated number of persons reached	Link to Zenodo upload
Organisation of a workshop	Nicola Botta (PIK), **TiPES Theme 4 Workshop**	Online, 13 October 2020	0	Scientific Community (higher education, Research)	40	
Organisation of a workshop	Nicola Botta (PIK), Nuria Brede (PIK), Michel Crucifix (UCL), Marina Martínez Montero (UCL), **Title: "Challenges and new directions in risk analysis, decision making and policy advice for climate change"**	Online, 11 March 2022	0	Scientific Community (higher education, Research)	40	Archived version of workshop webpage: https://web.archive.org/web/20220601114041/ Current web version: https://www.pik-potsdam.de/members/nubrede/tipes-wp6-workshop-challenges-and-new-directions-for-climate-change-related-risk-analysis-and-decision-making
other	[BBCM21a] Nicola Botta (PIK), Nuria Brede (PIK),		**0**	Scientific Community (higher	Circulated via TiPES mailing list,	https://doi.org/10.5281/zenodo.4543472

	Michel Crucifix (UCL), Cezar Ionescu (TH Deggendorf), Patrik Jansson (Chalmers UoT), Marina Martínez Montero (UCL), **Title: "A note on climate science and verified programming"**			education, Research)	publicly available online	
other	[BBCM21b] Nicola Botta (PIK), Nuria Brede (PIK), Michel Crucifix (UCL), Marina Martínez Montero (UCL), **Title: "A note on climate science and climate policy"**		**0**	Scientific Community (higher education, Research)	Circulated among colleagues, publicly available online	Zenodo: https://doi.org/10.5281/zenodo.6783575 EarthArXiv: https://doi.org/10.31223/X57M0S
other	[BBCM21c] Nicola Botta (PIK), Nuria Brede (PIK), Michel Crucifix (UCL), Marina Martínez Montero (UCL), **Title: "Decision Theory and Climate Policy"**		**0**	Scientific Community (higher education, Research)	prepared as input for WP7, publicly available online	https://doi.org/10.5281/zenodo.6783575
other	[Bot21] Nicola Botta (PIK), **Title: "The PI theorem and dimensional analysis"**		**0**	Scientific Community (higher education, Research)	circulated on the TiPES mailing list, publicly available online	https://doi.org/10.5281/zenodo.6783575
Participation to an event other than a conference or a workshop	Nuria Brede (PIK), **Title: "Types for TiPES - Applying Type Theory to Climate Impact Research",** PIK RD4 seminar	Potsdam, 11 February 2020	**0**	Scientific Community (higher education, Research)	~30	https://doi.org/10.5281/zenodo.4554685
Participation to a workshop	Nicola Botta (PIK), **Title: "The JFP2017 theory of verified policy advice: An overview"** TiPES Theme 4 Workshop	Online, 13 October 2020		Scientific Community (higher education, Research)	~40	https://doi.org/10.5281/zenodo.4545679
Participation to a workshop	Nuria Brede (PIK), Marina Martínez Montero (UCL),	Online, 13 October 2020		Scientific Community (higher	~40	https://doi.org/10.5281/zenodo.4554708

	Title: "Can Solar Radiation Management help to avoid Greenland's tipping point?" TiPES Theme 4 Workshop			education, Research)		
Participation to an event other than a conference or a workshop	Nicola Botta (PIK), **Title: "Bridging the gap between climate science and climate policy advice",** PIK Science & Pretzels	Online, 11 March 2020	**0**	Scientific Community (higher education, Research)	~30	https://doi.org/10.5281/zenodo.6783774
Participation to an event other than a conference or a workshop	Nicola Botta (PIK), **Title: "Climate science, program verification and policy advice", TiPES webinar**	Online, 04 November 2020	**0**	Scientific Community (higher education, Research)	~50	https://doi.org/10.5281/zenodo.4543621
Participation to an event other than a conference or a workshop	Nicola Botta (PIK), **Title: "Responsibility under uncertainty: which climate decisions matter most?",** PIK RD4 seminar	Online, 30 March 2021	**0**	Scientific Community (higher education, Research)	~30	https://doi.org/10.5281/zenodo.6826502
Participation to an event other than a conference or a workshop	Nuria Brede (PIK), **Title: "Toward a DSL for Sequential Decision Problems with Tipping Point Uncertainties",** PIK RD4 seminar	Online, 05 October 2021	**0**	Scientific Community (higher education, Research)	~30	https://doi.org/10.5281/zenodo.6783894
Participation to an event other than a conference or a workshop	Nicola Botta (PIK), **Title: "Equations, transformations and dimensions: better types?",** Chalmers Functional Programming Group Seminar	Online, 04 March 2022	**0**	Scientific Community (higher education, Research)	~30	https://doi.org/10.5281/zenodo.6783863
Participation to an event other than a conference or a workshop	Nicola Botta (PIK), **Title: "Equations, transformations and dimensions: better types?",** PIK RD4 seminar	Online, 03 May 2022	**0**	Scientific Community (higher education, Research)	~30	https://doi.org/10.5281/zenodo.6783800
Training	Nicola Botta (PIK), **Title: "Decision problems in climate research, mathematical specification and**	Online, 02+09 June 2022	**0**	Scientific Community (higher education, Research)	20	https://gitlab.pik-potsdam.de/botta/lectures/-/tree/master/2022.THD

	dependent types”, TH Deggendorf guest lectures					**Zenodo:** https://doi.org/10.5281/zenodo.6783894
Participation to an event other than a conference or a workshop	Nuria Brede (PIK), **Title: “On the correctness of monadic backward induction”,** PIK RD4 seminar	Potsdam, 14 June 2022	**0**	Scientific Community (higher education, Research)	5	https://doi.org/10.5281/zenodo.6783894

Peer reviewed articles

(Abbreviations: “Y” - “YES”, “P” - “published”, “SR - “submitted, under revision”)

Title	**Authors**	**Publication**	**DOI**	**Is TiPES correctly acknowledged?**	**How much did you pay for the publication?**	**Status?**	**Open Access granted**	**Comments on embargo time imposed by the publisher**	**If in Green OA, provide the link where this publication can be found**
Extensional equality preservation and verified generic programming	[Bot+21a] Nicola Botta (PIK), Nuria Brede (PIK), Patrik (Chalmers UoT) Jansson, and Tim Richter (U Potsdam).	Journal of Functional Programming	10.1017/S0956796821000204	Y	0	P	Y	0	https://doi.org/10.1017/S0956796821000204
On the correctness of monadic backward induction	[BB21] Nuria Brede (PIK), Nicola Botta (PIK)	Journal of Functional Programming	10.1017/S0956796821000228	Y	0	P	Y	0	https://doi.org/10.1017/S0956796821000228

(table continued on next page)

Responsibility under Uncertainty: Which Climate Decisions Matter Most?	[Bot+21b] Nicola Botta (PIK), Nuria Brede (PIK), Michel Crucifix (UCL), Cezar Ionescu (TH Deggendorf), Patrik Jansson (Chalmers UoT), Zheng Li (Northeastern University & Arima Inc.), Marina Martínez-Montero (UCL), Tim Richter (U Potsdam)	Environmental Modeling & Assessment	10.21203/rs.3.rs-1103231/v1	Y		SR	Y		https://doi.org/10.21203/rs.3.rs-1103231/v1

Other publications

Title	Authors	Type of document	Type of Audience	Link to Zenodo upload
Climate sensitivity, commitment and abrupt change: toward an ontology for climate tipping point research	[BBCM22a] Nuria Brede (PIK), Nicola Botta (PIK), Michel Crucifix (UCL), Marina Martínez Montero (UCL)	Report	Scientific Community (higher education, Research)	https://doi.org/10.5281/zenodo.6820683
A DSL for Monadic Decision Problems, Responsibility under Uncertainty and Tipping Point Notions	[BBCM22b] Nuria Brede (PIK), Nicola Botta (PIK), Michel Crucifix (UCL), Marina Martínez Montero (UCL)	Report	Scientific Community (higher education, Research)	https://doi.org/10.5281/zenodo.6820605
IdrisLibs, Version 2.0	[Bot22] Nicola Botta et al.	Software library	Scientific Community (higher education, Research)	https://doi.org/10.5281/zenodo.6822146

Literate Idris sources for [BBCM22b]	[BBCM22c] Nicola Botta (PIK), Nuria Brede (PIK), Michel Crucifix (UCL), Marina Martínez Montero (UCL)	Source code	Scientific Community (higher education, Research)	https://doi.org/10.5281/zenodo.6826927

Uptake by the targeted audiences

As indicated in the Description of the Action, the audience for this deliverable is:

X	The general public (PU) is and is made available to the world via CORDIS.
	The project partners, including the Commission services (PP)
	A group specified by the consortium, including the Commission services (RE)
	This report is confidential, only for members of the consortium, including the Commission services (CO)

How we are going to ensure the uptake of the deliverables by the targeted audiences

All the material presented in this deliverable is public, but the content is directed towards a scientific audience.

An important part of this deliverable, reported in Appendix B, has already been published in peer-reviewed journals or is currently under revision (papers [Bot+21a, Bot+21b, BB21]).

The work reported in Appendix A provides preliminary results for an ontology of operationalised key notions in the context of tipping point research and we plan to submit it as a contribution to EarthArxiv (open source) and advertise it through social media.

Since we are aware that the application of formal methods in climate science in general and to climate policy advice in particular is far from mainstream approaches, such as integrated assessment modelling or standard scenario simulation, we have assembled some thoughts about “climate science and verified programming”, “climate science and climate policy” and “decision theory and climate policy” in three short notes [BBCM21a-c], reported in Appendices C-E. These notes are intended to improve the visibility and understanding of the approach in the community. They have been made publicly available ([BBCM21a-c] on Zenodo and [BBCM21b] also on EarthArXiv), and have been announced on the TiPES Twitter account. The note [BBCM21c] was originally prepared for internal communication with TiPES WP7, and [BBCM21a, BBCM21b] have been circulated among external collaborators and on the TiPES mailing list. They have also been shared with the project managers of the tipping point related EU Horizon 2020 projects TiPACCS and COMFORT.

We have made all work reported in this deliverable publicly available on Zenodo or via other visible online platforms. This holds even for the source code from which the three papers [Bot+21a, Bot+21b, BB21] and the report [BBCM22b] in Appendix B have been created. Moreover, as a distinctive feature providing a high degree of trustworthiness, the formal mathematical content of these sources can be machine-checked for correctness by any interested member of the audience.

Appendices

Climate sensitivity, commitment and abrupt change: toward an ontology for climate tipping point research

Nuria Brede[1,2], Nicola Botta[1,3], Michel Crucifix[4], and Marina Martínez Montero[4]

[1]RD4: Complexity Science, Potsdam Institute for Climate Impact Research, Potsdam, Germany
[2]Department of Computer Science, University of Potsdam, Potsdam, Germany
[3]Department of Computer Science and Engineering, Chalmers University of Technology, Göteborg, Sweden
[4]Earth and Life Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium

{nuria.brede,botta}@pik-potsdam.de, {michel.crucifix,marina.martinez}@uclouvain.be

Abstract

We review a number of notions in tipping point research that have been studied within work package 6 of the TiPES project and sketch how to organise them in a prototype ontology. For *climate sensitivity* and *climate change commitment*, we propose an operational semantics by giving their generic computational structure.

Contents

1 Introduction

This short report is concerned with selected notions that are relevant in the context of *tipping point* research as conducted within the EU Horizon 2020 project TiPES (**Ti***pping* **P***oints in the Earth* **S***ystem*) [23]. The focus lies on the following notions:

- *Climate Sensitivity* (Section 4)
- *Commitment* (Section 5)
- *Abrupt climate change, Tipping Point (TP)* and *Tipping element (TE)* (Section 6)
- *Early Warning Signal (EWS)* (Section 7)

To explore these notions, we proceed as follows: We briefly give some general context (Section 2) and introduce a few abstract language elements to facilitate talking about models, simulations and data (Section 3). For each notion (Sections 4–7) we give a brief overview as introduction and collect (possibly informal) definitions and classification information that may be used as semantic annotation. We also assemble references for further reading that allow to follow the development of the notion in question. For climate sensitivity and commitment experiments, we suggest generic computational schemes of which these experiments are instances, following the approach of [Ion09; Ion16]. Our main motivation is ontological: How are the notions defined and used in the literature? What is their computational structure, which input and output types do they have, which physical dimensions (if applicable) and how are they related to other notions? In Section 8 we sketch an abstract perspective that may help to organise tipping point notions in the style of an ontology.

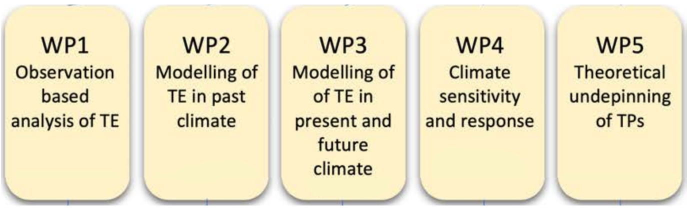

WP1
Observation based analysis of TE
WP2
Modelling of TE in past climate
WP3
Modelling of of TE in present and future climate
WP4
Climate sensitivity and response
WP5
Theoretical undepinning of TPs

Figure 1: TiPES work package topics

Remarks: The report is part of Deliverable 6.2 of TiPES WP6. In the context of WP6's objectives, it is meant to serve as preparation of a *domain specific language* (*DSL*) with tipping point notions. This DSL is to be implemented as extension of the *IdrisLibs* framework for the study of sequential decision problems [Bot21].[1]

[1] The framework is implemented in *Idris*, a dependently typed programming language that allows to specify, implement and prove properties of programs all in one language.

The choice of the notions discussed in this report is guided by the topics studied in TiPES WP1–5 (see Fig.1 and cf. Appendix II for reference). We include pointers in text passages that relate to these work packages' objectives (e.g. concerning the classical notion of *equilibrium climate sensitivity*). Parts of this work might also be understood as an elaboration of some entries of the *Intergovernmental Panel on Climate Change (IPCC)* glossary (its most recent version can be found in Appendix VII of the AR6 WGI report [Mas+21; Mat+21]) without the ambition of achieving the generality of the relevant chapters of the IPCC assessment reports. For reference, we have included a number of glossary entries which concern our notions of interest in Appendix I. Other general sources we found helpful in the preparation of this document are course materials/resulting textbooks [Goo+10; Goo15; Sto11] and the recent overview paper [GL20].

2 Preliminaries

Central notions underlying everything we will be discussing subsequently are those of *climate system* and *Earth system*.

2.1 Climate and Earth system

First of all: what is a *system*? Wikipedia tells us that

> "A system is a group of interacting or interrelated elements that act according to a set of rules to form a unified whole. [Mer] A system, surrounded and influenced by its environment, is described by its boundaries, structure and purpose and expressed in its functioning. Systems are the subjects of study of systems theory." *(citation adapted)*

and

> "In engineering and physics, a physical system is the portion of the universe that is being studied (of which a thermodynamic system is one major example). "

and also

> "Systems theory views the world as a complex system of interconnected parts. One scopes a system by defining its boundary; this means choosing which entities are inside the system and which are outside—part of the environment. One can make simplified representations (models) of the system in order to understand it and to predict or impact its future behavior. These models may define the structure and behavior of the system."

The TiPES project is concerned with tipping elements in the *Earth System* of which the *climate system* forms a subsystem. The *climate system* consists itself of five major subsystems, namely the atmosphere, hydrosphere, cryosphere, lithosphere and biosphere. The notion of *Earth system* in the context of *Earth System Science* [Com86; Com88; Mos06] is broader and in particular includes human societies as subsystems. The notions discussed in this document mostly have emerged from the study of the climate system, and the influence of human societies is usually considered as external *forcing* applied to the system, e.g. via anthropogenic CO_2 emissions. However, in Section 6 we will e.g. encounter the notion of *policy-relevant tipping element* which explicitly includes human *value judgements* and seeks to encompass tipping behaviour beyond bio-physical processes.
The IPCC glossary does not give a definition of *Earth System*, but a definition of *Climate System*.

IPCC Glossary: Climate system

The global system consisting of five major components: the atmosphere, the hydrosphere, the cryosphere, the lithosphere and the biosphere and the interactions between them. The climate system changes in time under the influence of its own internal dynamics and because of external forcings such as volcanic eruptions, solar variations, orbital forcing, and anthropogenic forcings such as the changing composition of the atmosphere and land-use change.

The IPCC considers the climate system as a dynamical system:

IPCC Glossary: Dynamical system

A process or set of processes whose evolution in time is governed by a set of deterministic physical laws. The climate system is a dynamical system.

Ghil and Lucarini begin their paper on the physics of climate change [GL20] with the statement:

> "The climate system is forced, dissipative, chaotic, and out of equilibrium; its complex natural variability arises from the interplay of positive and negative feedbacks, instabilities, and saturation mechanisms."

and later:

> "A key goal of climate modelling is to capture the system's statistical properties, i.e., its mean state and its variability, and its response to forcings of a different nature."

Accordingly, most of the notions we will discuss in this report are closely linked to studying the climate system as a dynamical system and its response to interventions of some form.

Because of the complexity of this system and the very limited possibility/ quasi impossibility to study it via systematic and repeatable empirical experiments, the main methodologies are based on physical theories and numerical simulations with *models.* There is a whole hierarchy of models, ranging from simple conceptual to complex global or regional models.

The study via models and model simulations is complemented with the analysis of data in the form of time series. The data either comes from *the instrumental record* [2] or from historical *proxies.* The outcome of assessments of climate system properties depends on the model and/or data used, and on additional assumptions that are made to fill gaps in the knowledge or simplifications necessary for computational feasibility.

Data sources can be classified according to the *time scales* for which they provide information. A rough orientation following Fig. 2 of [KR15]:

Timescales and different kinds of data

- Observations: years to decades
- GCM simulations: years to centuries (few to couple thousands of years)
- Paleo proxies: decades to hundreds of millions of years

Differences in the *time scale* of processes in the climate system can be formalised using the mathematical theory of *slow-fast systems* [Kue11].

2.2 Classification of climate models

Climate models can be classified according to features like which components they contain as submodels and the number of space dimensions. In more detail one can consider the components of the state space, the parameters, possible forcings and boundary conditions, but also aspects of computer implementations for numerical simulations. One speaks of a *model hierarchy* referring to the increase/decrease of complexity between different classes of models. The IPCC list on the lowest

[2] going back only a few hundred years, with systematic measurements starting in the second half of the 19th century

level of this hierarchy a class of *Simple Climate Models (SCMs)*, like *Energy Balance Models (EBMs)* or *ocean box models.* This class of models is used to study specific aspects of the climate system in a stylised way. On the highest level of the hierarchy one finds *Global Climate Models/General Circulation Models(GCMs)* and *Earth System Models (ESMs).* These models aim at a best possible representation of (bio)-physical processes. Integrating these models however comes with a high computational cost and thus it is not feasible to run huge ensembles of simulations for different scenarios. Another approach that seeks to combine advantages of simple and complex models is provided by *Earth system models of Intermediate Complexity (EMICs).*
Another way to distinguish different types of models is to classify them as *process-based, statistical* or *conceptual* models [Goo15; Cru12]. An overview over Earth system models is given in [Fla11].
An important example of the simplest class of models is the *Stommel box model* [Sto61] of the *thermohaline circulation (THC)* in the North Atlantic. It illustrates the northward transport at the surface of the ocean of relatively warm and salty waters which then cool down, sink and are transported back southward in the depth of the ocean. The THC is considered a tipping element and the Stommel model is an early example of a model with more than one stable state. The model and its variants have been and are still used in many studies (e.g. recently in [Alk+19; Loh+21; KLN22]).

Example 1. *(Stommel box model variant following [Mar00] and [Goo+10])*
The model consists of two "well-mixed boxes of equal volume" B_1 and B_2, containing water having a certain temperature and a certain salinity. Box B_1 represents the ocean at high latitudes and B_2 the ocean at lower latitudes. In the simplest variant of the model, the temperatures of the two boxes are not part of the dynamics but given as parameters.
The dynamic variables then are just $S_1, S_2 : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ representing the temporal evolution of the salinity of B_1 and B_2, respectively [3]
The model has the following parameters:

- $T_1, T_2 : \mathbb{R}$ – temperatures of boxes B_1 and B_2, respectively, with the condition that $T_2 > T_1$ (corresponding to lower temperatures at higher latitudes) ($[T_1] = [T_2] = \Theta$)[4]
- $k : \mathbb{R}_{>0}$ – a hydraulic constant ($[k] = \mathsf{L}^3\mathsf{M}^{-1}$)
- $\alpha : \mathbb{R}_{>0}$ – the thermal expansion coefficient ($[\alpha] = \mathsf{L}^{-3}\mathsf{M}\Theta^{-1}$)
- $\beta : \mathbb{R}_{>0}$ – the haline contraction coefficient ($[\beta] = \mathsf{L}^{-3}\mathsf{M}$)
- $H : \mathbb{R}$ – represents the surface salinity flux when positive and surface freshwater flux when negative (i.e. positive freshwater flux is presented as negative salinity flux) ($[H] = \mathsf{T}^{-1}\mathsf{L}^3$)

Two assumptions are made concerning the density of the water in the boxes and the flow between the boxes:

- the density of the water in each box can be approximated by a linear function $\rho : \mathbb{R}_{\geqslant 0} \times \mathbb{R} \to \mathbb{R}_{>0}$ of its salinity $s : \mathbb{R}_{\geqslant 0}$ and temperature $T : \mathbb{R}$

  $$\rho(s, T) = \rho_0 - \alpha * (T - T_0) + \beta * (s - s_0)$$

  where s_0, T_0 and ρ_0 are salinity, temperature and salinity at a reference state ($[\rho(s, T)] = \mathsf{L}^{-3}\mathsf{M}$)
- the strength of the salinity flow $q_{B_1 \to B_2} : \mathbb{R}$ between the two boxes (with densities $\rho_1, \rho_2 : \mathbb{R}_{>0}$, respectively) is proportional to their density difference:

  $$q_{B_1 \to B_2} = k * (\rho_1 - \rho_2) \tag{1}$$

[3] In an alternative formulation of the model there is only one dynamic variable representing the salinity difference between the two boxes.

[4] The bracket notation $[\cdot]$ is used to indicate physical dimension. Please see the paragraph "Physical dimensions" below for a discussion.

If $s_1, s_2 : \mathbb{R}_{\geqslant 0}$ are the salinities of the boxes, by combining these two assumptions one gets for $i \in \{1, 2\}$:

$$\rho_i = \rho(s_i, T_i) = \rho_0 - \alpha * (T_i - T_0) + \beta * (s_i - S_0)$$

and thus

$$\begin{aligned} \rho_1 - \rho_2 &= \rho_0 - \alpha * (T_1 - T_0) + \beta * (s_1 - S_0) - (\rho_0 - \alpha * (T_2 - T_0) + \beta * (s_2 - S_0)) \\ &= \alpha * (T_2 - T_1) + \beta(s_2 - s_1) \end{aligned}$$

Writing $\Delta T = T_2 - T_1$ for the difference between the box water temperatures that are given as parameters, and using the above assumptions, the salinity flow strength between the two boxes at some point in time $t : \mathbb{R}$ can then be computed via a function $q : \mathbb{R} \to \mathbb{R}$ given their salinity difference $\Delta S(t) = S_2(t) - S_1(t)$ at time t:

$$q(\Delta S(t)) = k * (\alpha * \Delta T - \beta * \Delta S(t))$$

If $q > 0$, the direction of the salinity flow is from low to high latitudes at the surface and from high to low at the bottom of the ocean (this corresponds to the current situation in the North Atlantic.) The state of the model represents the salinities of the two boxes. The equations that specify their temporal evolutions $S_1, S_2 : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ are:

$$\frac{dS_1}{dt}(t) = -H + |q(\Delta S(t))| * \Delta S(t) \tag{2}$$

$$\frac{dS_2}{dt}(t) = +H - |q(\Delta S(t))| * \Delta S(t) \tag{3}$$

where $\Delta S(t) = S_2(t) - S_1(t)$. That is, the change in salinity at time t depends on the surface salinity/freshwater flux and the product of the salinity flow strength between the two boxes and the salinity difference. Recall that box B_1 represents the cold water at high latitudes where the surface flux is a freshwater flux from meltwater (decreasing salinity). This is expressed in Eq. 2 as a negated salinity flux $-H$. For box B_2 the surface flux represents the evaporation at lower latitudes (increasing salinity). The flow strength is used as absolute value since the salinity balance of the boxes does not depend on whether the direction goes from B_1 to B_2 at the top and from B_2 to B_1 at the bottom or vice versa.

Assuming e.g. that the temperature difference is greater than the salinity difference between the two boxes ("temperature driven" situation), then, if the salinity of B_2 is higher than that of B_1, this decreases both the rate of change of the salinity of B_1 and B_2 and the flow strength. Otherwise both are increased. Whether the overall rate of change is positive or negative then depends critically on the value of the parameter H, as does the existence of an equilibrium solution (this will be relevant in Section 6).

The interest of the model is to illustrate what is called the *salinity* or *ocean advection feedback*: If the circulation strength of the *thermohaline circulation (THC)* (that is represented by the strength of the salinity flow in the model) is decreased by a perturbation, less salinity is transported to higher latitudes. This results in lower density there, which in turn decreases the flow strength thereby reinforcing the initial perturbation.

Classification of climate models

Some criteria that may be used to classify models:

- process-based, statistical, conceptual, ...
- model classes:
  - Energy Balance Model (*EBM*),
  - Earth system model of Intermediate Complexity (*EMIC*),

  - General Circulation/ Global Climate Models with representations of the atmosphere and/or the ocean (*AGCM, OGCM, AOGCM*)
  - Earth System Model (*ESM*)
  - Regional Climate Model (*RCM*),
  - Ocean box model
- spatial dimensions: 0, 1, 2, 3
- *subsystems* of the climate system represented in the model: atmosphere, ocean, ice-sheets, carbon cycle, ...

Physical dimensions. Following a notation originally introduced by Maxwell and widely applied in textbooks, we write $[e] = d$ to denote that the physical quantity e has dimension d. For example, we write $[T] = \Theta$ to indicate that T has the dimension of a temperature.
Expressions like $[e] = d$ and $[T] = \Theta$ are called *dimensional judgements*, in analogy with *type* judgements like $q \in \mathbb{R}$ or $q : \mathbb{R}$.
Specifically, $[T] = \Theta$ is an abbreviation for $[T] = \lambda(\mathsf{T}, \mathsf{L}, \mathsf{M}, \Theta).\Theta$ and the anonymous function[5] $\lambda(\mathsf{T}, \mathsf{L}, \mathsf{M}, \Theta).\Theta$ is called the *dimension function* of T. The judgement posits that measurements of T increase by Θ when the units of measurement for times, lengths, masses and temperatures are decreased by $\mathsf{T}, \mathsf{L}, \mathsf{M}, \Theta : \mathbb{R}_{>0}$, respectively, see [Bar+96] section 1.1.3.
Dimensional judgements are at the core of dimensional analysis and similarity theory [Buc14; Buc15; Bri22; Bar+96; Gib11] and build the basis of the mathematical modelling and of the data-based analysis of physical systems.

3 Framework

In this section we introduce a framework that we will use to describe computational structures in the remainder of the paper.

Model

A time-dependent climate model is given by a system of ODEs or PDEs:

$$\frac{\partial m}{\partial t} = \mathcal{F}(p, m). \tag{4}$$

where $\mathcal{F}$ is a higher-order function

$$\mathcal{F} : P \times (\mathcal{T} \times \mathbb{R}^d \to X) \to (\mathcal{T} \times \mathbb{R}^d \to X') \quad {}_{d \in \{0,1,2,3\}} \tag{5}$$

that, together with parameters $p : P$ and suitable initial and, possibly, boundary conditions, implicitly defines a function $m : \mathcal{T} \times \mathbb{R}^d \to X$.

Example 2. For an ODE of the form

$$\dot{m}(t) = f(p, t, m(t)) \tag{6}$$

with parameters $p : P$, time $t : \mathcal{T}$, functions $m : \mathcal{T} \to X$ and $f : P \times \mathcal{T} \times X \to X$ the higher order function $\mathcal{F}$ in the above framework is defined by

$$\mathcal{F}(p, m) = \lambda t. f(p, t, m(t)).$$

[5] Note that we use the λ-notation to define anonymous functions in the sense of Church's λ-calculus [Bar+84] as common in the context of functional programming but also e.g. in Python.)

Example 3. To express a PDE written

$$\frac{\partial m}{\partial t} = m * \frac{\partial m}{\partial r_1} \tag{7}$$

in abbreviation for

$$\frac{\partial m}{\partial t}(t,(r_1,r_2,r_3)) = m(t,(r_1,r_2,r_3)) * \frac{\partial m}{\partial r_1}(t,(r_1,r_2,r_3)) \tag{8}$$

with $t : \mathcal{T}$, $d = 3$,$(r_1, r_2, r_3) : \mathbb{R}^3$ and $m : \mathcal{T} \times \mathbb{R}^3 \to X$ in the above framework, we define

$$\mathcal{F}(p,m) = \lambda(t,(r_1,r_2,r_3)).m(t,(r_1,r_2,r_3)) * \frac{\partial m}{\partial r_1}(t,(r_1,r_2,r_3))$$

Example 4. *(continuing Example 1)*
For the Stommel model of Section 2, we can instantiate this scheme as follows:

- $\mathcal{T} = \mathbb{R}, d = 0$
- a state of the model represents the salinity of the two boxes at some point in time; the salinity cannot be negative: $X = \mathbb{R}^2_{\geqslant 0}$ [6]
- thus the sought solution m is of type $\mathbb{R} \to \mathbb{R}^2_{\geqslant 0}$
- the model is parameterised by constants $k, \alpha, \beta : \mathbb{R}_{>0}$ and $T_1, T_2, H : \mathbb{R}$, thus $P = \mathbb{R}^3_{>0} \times \mathbb{R}^3$. If we have a closer look at the defining equations, though, these parameters could actually be merged into just one.
- the strength of the flow is given by a function $q : P \times \mathbb{R} \to \mathbb{R}$ with
  $$q(p, \Delta s) = k * (\alpha * \Delta T - \beta * \Delta s)$$
  where $p = (k, \alpha, \beta, T_1, T_2, H)$ and $\Delta T = T_2 - T_1$.
- with $f : \mathbb{R}^3_{>0} \times \mathbb{R}^3 \times \mathbb{R} \times \mathbb{R}^2_{\geqslant 0} \to \mathbb{R}^2$
  $$f(p, t, S_1, S_2) = (-H + |q(p, \Delta S)| * \Delta S, +H - |q(p, \Delta S)| * \Delta S)$$
  where $\Delta S = S_2 - S_1$. Then $\mathcal{F}(p, m) = \lambda t. f(p, t, m(t))$ as above.

Dynamical Systems. As we have seen in the last section, to study its properties, the climate system is considered as a dynamical system. One may consider notions of dynamical systems of increasing complexity. We mention here *autonomous, nonautonomous* and *monadic* dynamical systems.
Following [Kuz13, Def.1.1]:

Autonomous Dynamical System

An *autonomous dynamical system* is a triple $(\mathcal{T}, X, \{\varphi^t\}_{t:\mathcal{T}})$, where $\mathcal{T}$ is a time set, X is a state space, and $\{\varphi^t : X \to X\}_{t:\mathcal{T}}$ is a family of evolution operators satisfying the following properties:

$$\varphi^0 = id \tag{9}$$

$$\varphi^{t_1+t_2} = \varphi^{t_1} \circ \varphi^{t_2} \tag{10}$$

for all $x : X,\ t_1, t_2 : \mathcal{T}$ such that both sides of the equations are defined when applied to x (and where id is the identity function on X).

For continuous dynamical systems, the family of evolution operators $\{\varphi^t\}_{t:\mathcal{T}}$ is called a *flow.* Note that $\varphi^t(x)$ is not necessarily defined for all combinations of $t : \mathcal{T}$ and $x : X$.
Similarly, for nonautonomous systems (cf. [KR11, Def.2.1]):

[6] sometimes the dynamics is simply expressed for the salinity difference, in which case we simply have $X = \mathbb{R}$

Nonautonomous Dynamical System

An *nonautonomous dynamical system* is a triple $(\mathcal{T}, X, \{\varphi^t\}_{t:\mathcal{T}})$, where $\mathcal{T}$ is a time set, X is a state space, and $\{\varphi^t : \mathcal{T} \times X \to X\}_{t:\mathcal{T}}$ is a family of evolution operators satisfying the following properties:

$$\varphi^0(t_0, \cdot) = id \qquad \forall t_0 \in \mathcal{T} \tag{11}$$

$$\varphi^{t_1+t_2}(t_0, \cdot) = \varphi^{t_2}(t_0 + t_1, \cdot) \circ \varphi^{t_1}(t_0, \cdot) \qquad \forall t_0 \leqslant t_1 \leqslant t_2 \in \mathcal{T} \tag{12}$$

for all $x : X, t_0, t_1, t_2 : \mathcal{T}$ such that both sides of the equations are defined when applied to x (and where id is the identity function on X).

Not all systems of PDEs or ODEs specify a dynamical system. But if for any initial time t_0 and initial state function $x_0 : \mathbb{R}^d \to X$ there is a unique solution x_{x_0} for Eq. 4 such that for all $r : \mathbb{R}^d$

$$x_{x_0}(t_0, r) = x_0(r) \tag{13}$$

then a nonautonomous dynamical system $(\mathcal{T}, \mathbb{R}^d \to X, \{\varphi^t\}_{t:\mathcal{T}})$ can be derived by defining

$$\varphi^t(t_0, x_0)(r) = x(t_0 + t, \cdot) = x_0(r) + \int_{t_0}^{t_0+t} \mathcal{F}(p, x)(\tau, r) d\tau. \tag{14}$$

To represent different kinds of non-determinism in dynamical systems, we use the abstract notion of *monad* [Mac78], following Ionescu [Ion09]. This approach leads to the notion of *monadic dynamical system*. Extending the above notion of nonautonomous dynamical system we define:

Nonautonomous Monadic Dynamical System

A *nonautonomous monadic dynamical system* is a quadruple $(\mathbf{M}, \mathcal{T}, X, \{\varphi^t_{\mathbf{M}}\}_{t:\mathcal{T}})$, where $\mathbf{M} = (M, \mu, \eta)$ is a monad, X is a state space, and $\{\varphi^t_{\mathbf{M}} : \mathcal{T} \times X \to M(X)\}_{t:\mathcal{T}}$ is a family of evolution operators satisfying the following properties:

$$\varphi^0_{\mathbf{M}}(t_0, \cdot) = \eta_X \qquad \forall t_0 \in \mathcal{T} \tag{15}$$

$$\varphi^{t_2}_{\mathbf{M}}(t_0, \cdot) = \varphi^{t_2}(t_1, \cdot) \circ_{\mathbf{M}} \varphi^{t_1}_{\mathbf{M}}(t_0, \cdot) \qquad \forall t_0 \leqslant t_1 \leqslant t_2 \in \mathcal{T} \tag{16}$$

where $\circ_{\mathbf{M}}$ denotes the composition of the Kleisli category of the monad $\mathbf{M}$.

The (families of) operations associated with a monad $\mathbf{M} = (M, \mu, \eta)$ have the signatures (for any type A)

$$\mu_A : M(M(A)) \to M(A)$$

$$\eta_A : M(M(A)) \to M(A)$$

We see that M is an operation that maps types into types:

$$M : Type \to Type$$

Moreover M needs to be a *functor* which means that there is for any types A, B an operation

$$map_{A,B} : (A \to B) \to (M(A) \to M(B))$$

that lifts functions into the monad. Subscripts for these functions are usually left implicit.
Examples of monad functors are the powerset functor $\mathcal{P}$ and the list functor *List* which can be used to model non-determinism, or a functor *Prob* of probability distributions [7]. The identity functor *Id* is also a monad and allows to recover deterministic dynamical systems from monadic ones.

[7] there are monads both for discrete and continuous probability distributions [Gir81; EK06; Jac15]

Example 5. The operations associated with the list monad are defined as follows:

$$\begin{aligned}
&\mathit{map}_{A,B} : (A \to B) \to (\mathit{List}(A) \to \mathit{List}(B)) && \\
&\mathit{map}_{A,B}(f)([\,]_A) &=\;& [\,]_B \\
&\mathit{map}_{A,B}(f)(a ::_A as) &=\;& (f(a) ::_B \mathit{map}_{A,B}(f)(as)) \\
\\
&\eta_A : A \to \mathit{List}(A) && \\
&\eta_A(a) &=\;& [a]_A \\
\\
&\mu_A : \mathit{List}(\mathit{List}(A)) \to \mathit{List}(A) && \\
&\mu_A([[\,]_A]_{\mathit{List}(A)}) &=\;& [\,]_A \\
&\mu_A((as ::_{\mathit{List}(A)} ass)) &=\;& as +\!\!+_A \mu_A(ass)
\end{aligned}$$

where $[\,]_A : \mathit{List}(A)$ and $(::)_A : A \times \mathit{List}(A) \to \mathit{List}(A)$ are the *constructors* of the list type, $(+\!\!+)_A : \mathit{List}(A) \times \mathit{List}(A) \to \mathit{List}(A)$ is the operation that appends two lists and $[a]_A$ is a notation for $(a ::_A [\,]_A)$. The binary operations $(::)$ and $+\!\!+$ are written infix.

When we subsequently use lists, subscripts will be dropped if they can be inferred from the context.

Time $\mathcal{T}$ and time series

- The time set of a dynamical system can be $\mathbb{R}$ (for *continuous* dynamical systems), $\mathbb{Z}, \mathbb{N}$ (for *discrete* dynamical systems) or more generally a non-empty set that at least carries the structure of a monoid [Kuz13; GM12].
- *Time* in physical models is *continuous*, i.e. $\mathcal{T} = \mathbb{R}$.
- The index set $\mathcal{I}$ of any sequence of observations (in their original form) is necessarily discrete and finite, corresponding to an index type $\mathcal{I} = \mathbb{N}_{<n}$ for some $n : \mathbb{N}$. To obtain a continuous sequence of data from observations of type D, a function $\mathbb{R} \to D$ has to be reconstructed from the given data points.
- A time series is called *regular* if no value is missing and values are equally spaced in time [Dak+12].
- Sequences of proxy data are typically not equally spaced, in the sense that the time spans between subsequent data points might vary. To make time series of observations from different sources comparable, sequences have to undergo a *dating* process to obtain a regular time series (see Ch.5.3.2 of [Goo15] for different dating methods).
- Computer simulations require some form of discretisation.

To bridge from the mathematical definition of a model and a continuous dynamical system to a computational framework, we need methods to obtain a discrete-time dynamical system that can approximate the dynamics described by a model or continuous-time system. Moreover such a computational framework should be flexible enough to vary model parameters and forcings. We do not dwell on discretisation or numerical methods in this report, we just note that given appropriate such methods, we can derive a discrete dynamical system roughly as follows:
Given a model with associated flow φ and a (finite) [8] time discretisation $ts : \mathbb{N}_{\leqslant N} \to \mathcal{T}, N : \mathbb{N}, N > 0$, we can define a one step function

$$\begin{aligned}
&\mathit{next} : \mathbb{N}_{<N} \times X \to X \\
&\mathit{next}(k, x) = \varphi^{ts(k+1) - ts(k)}(ts(k), x)
\end{aligned} \tag{17}$$

[8] One might alternatively want to determine a time step for arbitrary many steps via a function $\mathbb{N} \to \mathcal{T}$, which would allow to compute simulations without fixing the number of computation steps in advance until a termination criterion is met. Such computations are not in general guaranteed to terminate and we do not discuss them further in this document.

and its iteration

$$\begin{array}{lcl} \mathit{flow} : \mathbb{N}_{<N} \times X \to X & & \\ \mathit{flow}(0, x_0) & = & x_0 \\ \mathit{flow}(n+1, x_0) & = & \mathit{next}(n, \mathit{flow}(n, x_0)) \end{array} \tag{18}$$

to approximate φ on the interval $[ts(0), ts(N)]$ and compute trajectories by

$$\begin{array}{lcl} \mathit{trajectory} : \mathbb{N}_{<N} \times X \to \mathit{List}(X) & & \\ \mathit{trajectory}(0, x_0) & = & [x_0] \\ \mathit{trajectory}(n+1, x_0) & = & \tau \mathbin{+\!\!+} [\mathit{next}(n, \mathit{last}(\tau))] \\ & & \text{where } \tau = \mathit{trajectory}(n, x_0) \end{array} \tag{19}$$

where $+\!\!+$ appends lists and *last* returns the last element of a list.
Similar operations can be derived for a monadic dynamical system with flow $\varphi_{\mathbf{M}}$ for $\mathbf{M} = (M, \mu, \eta)$, using the monad's unit η and the functorial *map* of the underlying functor M.

$$\begin{array}{l} \mathit{next}_{\mathbf{M}} : \mathbb{N}_{<N} \times X \to M(X) \\ \mathit{next}_{\mathbf{M}}(k, x) = \varphi_{\mathbf{M}}^{ts(k+1)-ts(k)}(ts(k), x) \end{array} \tag{20}$$

$$\begin{array}{lcl} \mathit{flow}_{\mathbf{M}} : \mathbb{N}_{<N} \times X \to M(X) & & \\ \mathit{flow}_{\mathbf{M}}(0, x_0) & = & \eta_X(x_0) \\ \mathit{flow}_{\mathbf{M}}(n+1, x_0) & = & \mathit{map}(\lambda x.\mathit{next}(n, x))(\mathit{flow}(n, x_0)) \end{array} \tag{21}$$

$$\begin{array}{lcl} \mathit{trajectory}_{\mathbf{M}} : \mathbb{N}_{<N} \times X \to M(\mathit{List}(X)) & & \\ \mathit{trajectory}_{\mathbf{M}}(0, x_0) & = & \eta_{\mathit{List}(X)}([x_0]) \\ \mathit{trajectory}_{\mathbf{M}}(n+1, x_0) & = & \mathit{map}(\lambda xs.xs \mathbin{+\!\!+} [\mathit{next}(n, \mathit{last}(xs))])(m\tau) \\ & & \text{where } m\tau = \mathit{trajectory}_{\mathbf{M}}(n, x_0) \end{array} \tag{22}$$

such that $(\mathbf{M}, \mathbb{N}, X, \{\mathit{flow}^{\mathbf{M}}(n, \cdot, \cdot)\}_{n:\mathbb{N}})$ is a discrete monadic dynamical system.

For numerical simulations, the deterministic *trajectory* function allows us to compute n-step approximations $[x_0, \ldots, x_n] : \mathit{List}(X)$. [9] Similarly, the monadic version of $\mathit{trajectory}_{\mathbf{M}}$ computes monadic values $mxs : M(\mathit{List}(X))$ containing such n-step trajectories. E.g. for a probability monad one would thus obtain a probability distribution of trajectories, and for the power set monad a set of trajectories.

In the following sections we discuss the key notions listed in the introduction using the framework introduced above.

4 Climate Sensitivity

Climate Sensitivity (CS) is a measure of the change of the *global mean surface temperature* (*GMST*) in response to the change of the CO_2 concentration in the atmosphere. The first estimates have been published as early as 1896 [Arr96].
Since the 1960s scientists have been systematically studying the GMST change in response to the change of CO_2 content in the atmosphere along what is referred to as different lines of evidence [KRH17], combining model simulations with models of different levels of complexity with instrumental or paleoclimatic proxy data.
The classic *equilibrium climate sensitivity* (*ECS*) experiment estimates the change in GMST relative to a reference state when the climate system will have settled to a new equilibrium after having been perturbed by doubling the CO_2 concentration in the atmosphere.

[9] One might additionally want to indicate the order of the approximation, e.g. for a 0-dimensional model with solution $x : \mathcal{T} \to X$, a q-order approximation would guarantee that $x_k = x(t_k) + \mathcal{O}((\Delta t)^q)$ for a step length of $\Delta t : \mathcal{T}$.

Climate sensitivity, defined as a change in temperature measured between two equilibrium states, is and will remain a theoretical construction. The explicit measure of the temperature change at equilibrium cannot be made, in practice or in principle, because it would require "fixing" both external factors (i.e., the orbital forcing) and even internal components of the climate system (e.g., the "ice sheets") and controlling accurately the CO_2 concentration.

Such experiments can however be made with climate models (simple and more complex), and it is our understanding of the connection between these models and the real world that allows us to attribute and estimate a "climate sensitivity" of our Earth. The meaning and intuition about climate sensitivity as a change in "equilibrium" states generally involves assumptions of time scale separation: climate sensitivity is understood as the measured response of some "fast" components of the climate system (e.g., the atmosphere) assuming that other components of the climate system have not changed (e.g., the ice sheets). As we will develop below, different assumptions about what is "fast" and "slow" lead to different definitions of the climate sensitivity, and [Roh+12] introduces a notation to that end.

The time scale separation is an arguable assumption (e.g., the adjustment time scale for the deep ocean temperature is of the order of 5000 years, a time scale over which the orbital forcing is significantly changed). Definitions of climate sensitivity that do not necessitate time scale separation have been proposed (e.g., [Ghi14] defines it as the growth of the Wasserstein distance between two pullback attractors) but their meaning is arguably much less intuitive.

Definitions. In the IPCC assessment reports, *equilibrium climate sensitivity* is defined as the change of GMST for the radiative forcing resulting from a doubling of the CO_2 level in the atmosphere wrt pre-industrial[10], a temperature difference (see p. TS-14, [Mas+21]).

Another approach to ECS is based on an *equilibrium climate sensitivity parameter* [Roh+12] which quantifies a change in GMST per unit change in radiative forcing.

Equilibrium climate sensitivity (ECS)

ECS: The *ECS* quantifies the change of GMST for the radiative forcing resulting from an intervention which doubles the CO_2 level in the atmosphere wrt preindustrial

$$\Delta T_{2\times \mathrm{CO_2}} = T_{\mathrm{eq}} - T_{\mathrm{pi}} \qquad \Delta T_{2\times \mathrm{CO_2}}, T_{\mathrm{eq}}, T_{\mathrm{pi}} : \mathbb{R} \qquad [\Delta T_{2\times \mathrm{CO_2}}] = [T_{\mathrm{eq}}] = [T_{\mathrm{pi}}] = \Theta$$

where T_{pi} is the GMST at the preindustrial reference state and T_{eq} the equilibrium GMST reached when the system has returned to radiative equilibrium after the intervention.

ECS Parameter: The *ECS parameter* $S_\bullet$ quantifies the change in mean surface temperature per unit change in radiative forcing

$$S_\bullet : \mathbb{R} \qquad [S_\bullet] = \Theta \mathsf{M}^{-1}\mathsf{T}^{3}.$$

According to [Roh+12], it is typically assumed that $\Delta T_{2\times \mathrm{CO_2}}$ is related to $S_\bullet$ by

$$\Delta T_{2\times \mathrm{CO_2}} = S_\bullet * \Delta Q_{2\times \mathrm{CO_2}} \tag{23}$$

where $\Delta Q_{2\times \mathrm{CO_2}} : \mathbb{R}, [\Delta Q_{2\times \mathrm{CO_2}}] = \mathsf{MT}^{-3}$ is the difference between the amount of radiative forcing resulting from a doubling of the CO_2 level in the atmosphere and a reference value prior to the doubling intervention.

According to Goosse [Goo15] "relatively good approximations" of the change in radiative forcing resulting from a change in CO_2 concentration "can be obtained [...] from a simple formula". Myhre et al. [Myh+98] give the following formula[11]:

$$\Delta Q_{\mathrm{CO_2}} : \mathbb{R}^2 \to \mathbb{R} \qquad \Delta Q_{\mathrm{CO_2}}(C, C_{\mathrm{ref}}) = 5.35 * \ln\left(\frac{C}{C_{\mathrm{ref}}}\right) \mathrm{W\,m^{-2}} \tag{24}$$

[10] since the pre-industrial state is assumed to be in radiative equilibrium

[11] Myhre et al. [Myh+98] also give formulas for other greenhouse gases, e.g. CH_4 and N_2O.

where $C, C_{\mathrm{ref}} : \mathbb{R}$ are CO_2 concentrations of the atmosphere (dimensionless), for the current and a reference state, respectively, and $[\Delta Q_{\mathrm{CO}_2}(C, C_{\mathrm{ref}})] = \mathsf{MT}^{-3}$. Using this formula, a doubling of the CO_2 concentration of the atmosphere results in a change of radiative forcing of $5.36 * \ln 2 \approx 3.7\mathrm{W\,m^{-2}}$.

The informal definitions given above suggest the following formalisation:
Let the evolution of the climate system with state space X be described by the flow $\{\varphi^t : X \to X\}_{t \in \mathcal{T}}$ and $x_{\mathrm{pi}} : X$ be the pre-industrial state of the climate. Let moreover *double* $: X \to X$ be a function that doubles the CO_2 content of the atmosphere and *gmst*, $C_{\mathrm{CO}_2} : X \to \mathbb{R}$ be functions that compute the GMST and the CO_2 concentration in the atmosphere, respectively. Then

$$\Delta T_{2\times\mathrm{CO}_2} = T_{\mathrm{eq}} - T_{\mathrm{pi}}$$

where

$$T_{\mathrm{eq}} = \mathit{gmst}\left(\lim_{t\to\infty} \varphi^t(\mathit{double}(x_{\mathrm{pi}}))\right) \qquad \text{and} \qquad T_{\mathrm{pi}} = \mathit{gmst}(x_{\mathrm{pi}})$$

and, if $S_\bullet$ is defined based on $\Delta T_{2\times\mathrm{CO}_2}$,

$$S_\bullet = \frac{\Delta T_{2\times\mathrm{CO}_2}}{\Delta Q_{2\times\mathrm{CO}_2}} \approx 0.3 * \Delta T_{2\times\mathrm{CO}_2}$$

$$\text{where} \qquad \Delta Q_{2\times\mathrm{CO}_2} = \Delta Q_{\mathrm{CO}_2}(C_{\mathrm{CO}_2}(\mathit{double}(x_{\mathrm{pi}})), C_{\mathrm{CO}_2}(x_{\mathrm{pi}})) \approx 3.7\mathrm{W\,m^{-2}}$$

ECS as a standard metric for climate models is however more specific with respect to the processes that are modelled (see below). One might also want to represent internal variability of the climate system or uncertainty about the pre-industrial climate state by using a stochastic model.

Radiation balance. The relation between $\Delta T_{2\times\mathrm{CO}_2}$ and $S_\bullet$ given in Eq. (23) follows from the idea [Goo15; Han+84] that, given a particular change in radiative forcing $\Delta Q : \mathbb{R}$, the radiative balance (the difference between incoming and outgoing radiation at the top of the atmosphere) $\Delta R(x) : \mathbb{R}, [\Delta R(x)] = \mathsf{MT}^{-3}$ for a climate state $x : X$ can be estimated by a linear function $\tilde{R} : \mathbb{R} \to \mathbb{R}$ with

$$\tilde{R}(\Delta T(x)) = \Delta Q - \tfrac{1}{S_\bullet} * \Delta T(x) \tag{25}$$

which takes a temperature difference $\Delta T(x) = \mathit{gmst}(x) - \mathit{gmst}(x_{\mathrm{ref}})$ between the climate state and a reference state $x_{\mathrm{ref}} : X$ as input.
The system is in radiative equilibrium in state x if $\Delta R(x) = 0$. This is the case if

$$\Delta T(x) = \Delta Q * S_\bullet \tag{26}$$

Doing this calculation for $\Delta Q_{2\times\mathrm{CO}_2}$, we get (an estimation of) $\Delta T_{2\times\mathrm{CO}_2}$ as in Eq. (23) above.
On the other hand, if $\Delta Q_{2\times\mathrm{CO}_2}$ and $\Delta T_{2\times\mathrm{CO}_2}$ have been determined by a numerical simulation, Eq. 26 can be used to calculate (an estimation of) a value for $S_\bullet$.

The negated inverse of the ECS parameter, $-S_\bullet^{-1}$ (the slope of $\tilde{R}$ above) is called the *climate feedback*. It can in turn be understood as being the sum of multiple different feedbacks (see below).

Variants. Estimates of ECS based on different data sources vary. Reasons for these differences are recently being addressed in the literature, and also other variants of climate sensitivity have been proposed. The IPCC lists the following variants of climate sensitivity:

Climate Sensitivity: Variants

- *equilibrium climate sensitivity (ECS)*
- *effective climate sensitivity (EECS)*

- *earth system sensitivity (ESS)*
- *transient climate response (TCR)*
- *transient climate response to cumulative CO_2 emissions (TCRE)*

Using the notion of radiative forcing, the IPCC calls different variants of climate sensitivity *climate metrics*: "Measures of aspects of the overall climate system response to radiative forcing" [Mat+21]. Given the IPCC definition of ECS as response to a change in the CO_2 concentration, this definition of *climate metric* seems to implicitly translate a change in CO_2 concentration into an estimation of corresponding radiative forcing.
We will address the differences between these variants in Section 4.1 below.

Estimation methods. Knutti et al. [KRH17] write about different ways of estimating climate sensitivity (and give an extensive overview over different estimates in the literature in their figures 1–3):

> "ECS and TCR cannot be measured directly, but in principle they can be estimated from:
>
> (i) quantifying feedbacks, ECS and TCR in comprehensive climate models;
>
> (ii) potentially constraining models by their representation of present-day mean climate and variability;
>
> (iii) analysis of the post-industrial observed warming of the ocean and atmosphere in response to forcing
>
> (iv) the short-term climate response to forcing (such as volcanic eruptions) or interannual temperature variations;
>
> (v) paleoclimate records (for example, the cooling at the Last Glacial Maximum or the warming during earlier warm periods). "

The authors of [Roh+12] emphasise that in order to make estimates of climate sensitivity comparable, it is important to be transparent about the *climate feedbacks* a particular calculation accounts for. Knutti and Rugenstein illustrate which feedbacks are typically included in the different variants of *CS*, on which time scales these feedbacks occur (in Fig. 1(b) and Fig. 2, respectively, of [KR15]) and also indicate which time scales relate to which kind of data (model simulation output, instrumental and proxy records). However, they also remark:

> "The separation of ECS and ESS is often made along timescales, with the argument that those feedbacks included in ECS essentially scale with surface temperature, whereas others in ESS partly have their intrinsic (and often slower) timescales. However, this does not apply to atmospheric chemistry which responds quickly. Here, the reason is a historic one, as the early climate models simply did not simulate interactive chemistry. This supports the argument that the separation of ECS and ESS is somewhat arbitrary in the real world where a lot of processes interact."

Climate feedbacks: Examples

The estimations of climate sensitivity differ in terms of the climate feedbacks taken into account

- *ECS*, *TCR*, *TCRE* include the following up to centennial scale ("fast") feedbacks:
  - clouds
  - lapse rate
  - water vapour

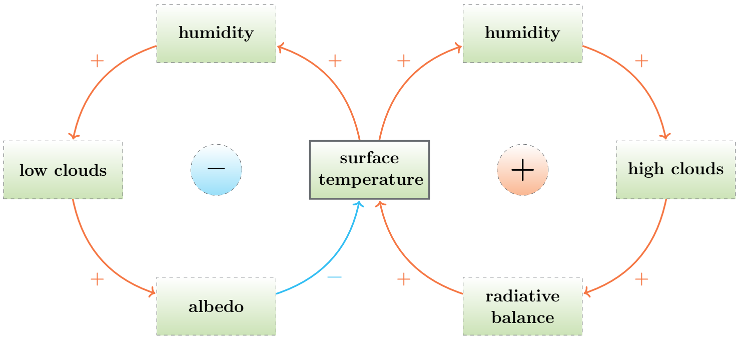


Figure 2: Example of positive and negative feedback loops. (Redrawn from TiPES Deliverable D4.1)

  - albedo/land surface
  - Planck

- *ESS*: Additionally included feedback processes:
  - atmospheric chemistry (another "fast" feedback that however was not available in earlier generations of process-based models)
  - up to millennia or longer ("slow"):
    - ocean circulation, ocean chemistry, weathering
    - dynamic vegetation, terrestrial ecosystems
    - permafrost carbon
    - ice sheets

Which of these feedbacks contribute to the calculation of a particular variant of CS is determined by the model that is being used. The ECS parameter can be refined by considering a sum of feedback parameters

$$\Sigma_{i=1}^{n}\lambda_i \qquad \lambda_i : \mathbb{R} \qquad [\lambda_i] = \Theta^{-1}\mathsf{MT}^{-3} \qquad 1 \leqslant i \leqslant n : \mathbb{N} \tag{27}$$

and defining the ECS parameter as (cf. [KR15; Roh+12])

$$S_\bullet = -\frac{1}{\Sigma_{i=1}^{n}\lambda_i}. \tag{28}$$

4.1 Computational structure of *CS* experiments

We have to distinguish between simple EBMs in which climate sensitivity is included as a parameter (or based on a small number of feedback parameters) and more complex models in which climate sensitivity is an "emergent property" and measured as response to perturbation of the system.
As seen above, different methods to estimate climate sensitivity are used in the literature. Here we describe two representative ones:

1. performing a perturbation experiment with a complex climate model to observe the resulting change in temperature (as emergent behaviour);

2. given observation or proxy data for a reference period, use a simple model which is parameterised over $S_\bullet$, perform model simulations for that reference period using different plausible values $S_\bullet$, and compare the model output to the given data to see which choice of the parameter fits best according to a chosen measure.

Estimation by perturbation experiment. The first approach consists in a perturbation experiment with a complex model in which climate sensitivity is an emergent property. We describe this method in parallel for the different variants of CS.

Given:

- A model m with time set $\mathcal{T}$, state space X, parameters $(f, p') : (\mathcal{T} \to P) \times P'$, where the first component denotes a potentially time-dependent forcing, and a dynamical system induced by the model.
- An initial state $x_0 : X$ at time $t_0 : \mathcal{T}$, often a preindustrial state, in (near) radiative equilibrium or, a probability distribution on initial states $px_0 : \mathit{Prob}(X)$
- A perturbed initial state $x'_0 : X$, for example obtained by doubling the CO_2 concentration in the atmosphere of the initial state through a function $\mathit{double} : X \to X$, with $x'_0 = \mathit{double}(x_0)$ and $px'_0 = \mathit{map}(\mathit{double})(px_0) : \mathit{Prob}(\mathit{List}(X))$ in the probabilistic case.

Now the basic computation follows the scheme of what we might call a *comparison experiment*:

- Compute two numerical approximations for a time discretisation Δt

  $$xs_{\mathrm{ref}} = [x_0, \ldots x_n] = \mathit{trajectory}(n, x_0) \text{ and } xs = [x'_0, \ldots x'_n] = \mathit{trajectory}(n, x'_0)$$

  in the deterministic case or otherwise two probability distributions $pxs, pxs' : \mathit{Prob}(\mathit{List}(X))$ either by

  $$pxs_{\mathrm{ref}} = trajectory_{\mathit{Prob}}(n, x_0) \qquad pxs = trajectory_{\mathit{Prob}}(n, x'_0)$$

  or

  $$pxs_{\mathrm{ref}} = \mathit{map}(\mathit{trajectory}(n, \cdot))(px_0) \qquad pxs = \mathit{map}(\mathit{trajectory}(n, \cdot))(px'_0)$$

  xs_{ref} and pxs_{ref} play the role of reference trajectory and reference distribution of trajectories, respectively.

- Compare xs_{ref} and xs or pxs_{ref} and pxs, respectively, using functions

  $$\mathit{compare}_{\mathrm{det}} : \mathit{List}(X)^2 \to \mathit{Val}_c$$

  or

  $$\mathit{compare}_{\mathrm{prob}} : Prob(\mathit{List}(X))^2 \to \mathit{Val}_c$$

  where Val_c is a type of values, e.g. $\mathbb{R}$.

This comparison might for example be done in terms of the difference between the global mean surface temperature at statistically measured end states $x_f, x_{f_{\mathrm{ref}}} : X$ [12] obtained by applying a function

$$\mathit{measureX} : \mathit{List}(X) \to X$$

[12] Again, if we assume some form of time scale separation, the "statistical measure" involves that states x_f and x_{ref} are observed over a time long enough to effectively sample the fluctuations of the fast variables. In technical terms, we want to measure the invariant set of meteorological variables associated with the climate states x_f. Discarding this time-scale separation assumption comes at a huge cost: it involves introducing the notion of *pullback attractor* (e.g. [Ghi14]) that would be overwhelming for our purpose.

to xs and xs_{ref}:

$$gmst(x_f) - gmst(x_{f_{\text{ref}}}) \qquad \text{where} \qquad x_f = measureX(xs), \quad x_{f_{\text{ref}}} = measureX(xs_{\text{ref}}).$$

See below for remarks concerning different algorithmic possibilities for the comparison of trajectories.

The different variants of climate sensitivity now correspond to variations in the models and different ways of computing the numerical approximations (we just give the deterministic variants, the probabilistic/monadic ones can be obtained as above using *map*):

- - *ECS*: Choose a model m_{ECS} that represents the processes required to account for the climate feedbacks usually associated with ECS and a forcing

    $$Const_{\text{CO}_2} : \mathcal{T} \to \mathbb{R}$$

    that keeps the atmospheric concentration of CO_2 constant. Perturb initial state by doubling the CO_2 concentration and integrate the model maintaining this perturbation until an/a (approximate/statistical) radiative equilibrium state is reached after $n_{ECS} : \mathbb{N}$ steps
    The computation thus amounts to

    $$xs = (trajectory_{m_{ECS}}(n_{ECS}, \cdot) \circ double)(x_0)$$

    For the reference path integrate the model with the unperturbed initial state.

    $$xs_{\text{ref}} = trajectory_{m_{ECS}}(n_{ECS}, \cdot)(x_0)$$

    Then $\Delta T_{2\times\text{CO}_2}$ can be computed as indicated above by comparing the global mean surface temperature for the representative final states resulting from these two computations.
  - *ESS*: As for ECS but with a model m_{ESS} that accounts for all the feedbacks to be considered for ESS and a number of computation steps $n_{ESS} : \mathbb{N}$ such that all processes can (approximately) equilibrate.
  - *EECS*: Use a model m_{ECS} as for ECS. Perturb the initial state and integrate the model for $n_{EECS} : \mathbb{N}$ time steps representing $\approx 100 - 200$ yrs. Then estimate an equilibrium state $x_{\text{eq}} : X$ using a function $estimate_{\text{eq}} : List(X) \to X$[13]

    $$x_{\text{eq}} = (estimate_{\text{eq}} \circ trajectory_{m_{ECS}}(n_{EECS}, \cdot) \circ double)(x_0)$$

    Do the same with the unperturbed initial state to obtain a reference equilibrium state $x_{\text{ref-eq}}$. Compute the difference in GMST between the two equilibrium states as above.
  - *TCR*: Use a model m_{TCR} with a forcing

    $$Inc_{1\%\text{CO}_2} : \mathcal{T} \to \mathbb{R}$$

    representing a 1% increase of atmospheric CO_2 concentration per year until the double of the initial CO_2 concentration is reached.[14] (Again, the forcing is given as part of the model parameters, say $p = (Inc_{1\%\text{CO}_2}, p')$ where p' might be other parameters.) Integrate the model for $n_{TCR} : \mathbb{N}$ time steps, representing $\approx 60 - 80$ yrs. Then compute a representative climate state by averaging globally in space, and over 20 yrs in time, centred at the time of CO_2 doubling, using a function $climate : List(X) \to X$

    $$x_f = (climate \circ trajectory_{m_{TCR}}(n_{TCR}, \cdot))(x_0)$$

    Compute the GMST difference between x_f and a reference state.

[13] standard methods applied in studies cited by the IPCC are from [Han+05] – using fixed sea surface temperatures– and [Gre+04] – using linear regression

[14] This seems to require knowledge about the *CO* concentration for the initial state to define the forcing.

For a given dynamical system with state space X and a numerical method for computing $N : \mathbb{N}$ time steps, we suggest the following generic schemes:

$$\begin{aligned} & \mathit{comparisonExperiment} : (\mathit{List}(X)^2 \to \mathit{Val}_c) \times \mathbb{N}_{<N} \times X^2 \to \mathit{Val}_c \\ & \mathit{comparisonExperiment}\,(\mathit{compare}, n, x, x') \\ & \qquad = \quad (\mathit{compare} \circ \mathit{map}(\mathit{trajectory}(n, .)))(x, x') \end{aligned} \tag{29}$$

$$\begin{aligned} & \mathit{perturbExperiment} : (\mathit{List}(X)^2 \to \mathit{Val}_c) \times (X \to X) \times \mathbb{N}_{<N} \times X \to \mathit{Val}_c \\ & \mathit{perturbExperiment}\,(\mathit{compare}, \mathit{perturb}, n, x) \\ & \qquad = \quad \mathit{comparisonExperiment}(\mathit{compare}, n, x, \mathit{perturb}(x)) \end{aligned} \tag{30}$$

Based on these, one might describe a general CO_2 doubling CS experiment as

$$\begin{aligned} & \mathit{csExperiment} : \mathbb{N}_{<N} \times X \to \mathit{Val}_c \\ & \mathit{csExperiment}\,(n, x) \quad = \quad \mathit{perturbExperiment}(\mathit{gmst} \circ \mathit{measureX}, \mathit{double}, n, x) \end{aligned} \tag{31}$$

When such simulations are performed with multiple models, the results might in the simplest case be averaged. However, one might also incorporate information from model validations: the better a model is capable of reproducing observational data for a reference time period, the more credible it is. When aggregating CS values resulting from the same experiment with different models (as in the CMIP project), the individual values might thus be weighted according to the performance of the respective model in validation experiments.

Parameter experiment. The idea of validating the model output against reference data also forms the basis for the second method, yet in a different way.
This time, CS is not estimated as an emergent property of the model, but included in the model as a parameter.
This time we have

- A list of candidate values for the ECS parameter $cs = [S_0, \ldots, S_k]$.
- A list of models $[m_0, \ldots, m_k]$ with state space X and parameter space $P = \mathbb{R} \times P'$, where the first component of a parameter fixes the value of the ECS parameter such that model m_i has parameters (S_i, p') for $0 \leqslant i \leqslant k$.
- A time series $o : \mathcal{T} \to \mathit{Val}_o$ of observational data and a time discretisation $[t_0, \ldots, t_n]$ for an interval of interest.
- An initial state $x_0 : X$ compatible with the observations $o(t_0)$

With these inputs:

- Compute numerical approximations

  $$xs_i = [x_{i0}, \ldots x_{in}] = \mathit{trajectory}_{m_i}(n, x_0) \qquad (0 \leqslant i \leqslant k)$$

  for the different candidate values of the ECS parameter.
- Extract the observations of interest from the states in each xs_i via by mapping an observation function $\mathit{observe} : X \to \mathit{Val}_o$

  $$oxs_i = \mathit{map}(\mathit{observe})(xs_i)$$

  and compare the oxs_i to the observational reference data $os = [o(t_0), \ldots, o(t_n)]$ according to a distance metric $\mathit{compare} : \mathit{List}(\mathit{Val}_o)^2 \to \mathit{Val}_c$, resulting in a list of distances $[d_0, \ldots, d_k]$ where $d_i = \mathit{compare}(oxs_i, os)$.

- Determine which of the xs_i is the best approximation to the observational data using a function $\mathit{best} : (ds : \mathit{List}(\mathbb{R})) \to \mathbb{N}_{<length(ds)}$ which returns the index j of the best approximation.
- Return the candidate value for the ECS parameter that best fits the given observations os:[15]

  $$S_{\text{best}} = \mathit{nth}(cs, \mathit{best}(\mathit{map}(\mathit{compare} \circ \mathit{map}(\mathit{observe}))([xs_0, \ldots, xs_k])))$$

  where the auxiliary function $\mathit{nth} : (l : \mathit{List}(A)) \times \mathbb{N}_{<length(l)} \to A$ given a list and an index (smaller than the length of the list) returns the list element at that index.

Remarks. Above we did not address algorithmic particularities of *how* to compare trajectories. In fact, there are structurally different ways to proceed.
Given an observable

$$\mathit{observe} : X \to \mathit{Val}_o,$$

a comparison function

$$\mathit{compare} : \mathit{Val}_o^2 \to \mathit{Val}_c,$$

and an aggregation function for Val_c [16]

$$\mathit{aggregateVal}_c : \mathit{List}(\mathit{Val}_c) \to \mathit{Val}_c$$

one could first compute a point-wise comparison between the two trajectories and the aggregate the outcomes:[17] [18]

$$(\mathit{aggregateVal}_c \circ \mathit{map}^{List}(\mathit{compare}))(\mathit{zip}(\mathit{map}^{\blacktriangle}(\mathit{map}^{List}(\mathit{observe}))(xs, xs')))$$

spelled out: given a pair of trajectories $(xs, xs') : \mathit{List}(X)^2$, the function *observe* is mapped onto each of the two trajectories using map^{List} and $\mathit{map}^{\blacktriangle}$. This results in a pair of two lists of observations ($\mathit{List}(\mathit{Val}_o)^2$) which are “zipped” together resulting in a list of pairs of observations ($\mathit{List}(\mathit{Val}_o^2)$). Then the function *compare* is mapped onto this list of observation pairs, again using map^{List}. The resulting list of comparison values ($\mathit{List}(\mathit{Val}_c)$) is finally aggregated to a single outcome of type Val_c. [19]

Alternatively, given an aggregation function

$$\mathit{aggregateVal}_o : \mathit{List}(\mathit{Val}_o) \to \mathit{Val}_o$$

one may first aggregate the state observations and then compare the results:

$$(\mathit{compare} \circ \mathit{map}^{\blacktriangle}(\mathit{aggregateVal}_o \circ \mathit{map}^{List}(\mathit{observe})))(xs, xs')$$

The two options will not necessarily give the same results.
For the probabilistic case (and similar for an arbitrary monad), additionally a measure

$$\mathit{measureVal}_o : \mathit{Prob}(\mathit{Val}_o) \to \mathit{Val}_o,$$

or

$$\mathit{measureVal}_c : \mathit{Prob}(\mathit{Val}_c) \to \mathit{Val}_c$$

[15] Mapping a function f onto a list $[x_0, \ldots, x_n]$ means applying this function to each element of the list without changing the order of the elements, i.e. $\mathit{map}(f)([x_0, \ldots, x_n] = [f(x_0), \ldots, f(x_n)])$.

[16] which in turn might be induced by a binary operation $\oplus : \mathit{Val}_c^2 \to \mathit{Val}^c$

[17] Below we annotate the *map* function for the different functors for better readability. We write $\mathit{map}^{\blacktriangle}$ for the map of the diagonal functor $\mathit{Diag} : \mathit{Type} \to \mathit{Type}$ with $\mathit{Diag}(A) = A \times A$, map^{List} for the map of the list functor defined in Example 5 and map^{Prob} for the map of a functor that maps each type to the type of probability distributions over this type.

[18] The function $\mathit{zip} : \mathit{List}(A) \times \mathit{List}(B) \to \mathit{List}(A \times B)$ transforms a pair of lists of equal lengths into a list of pairs of the same length: $\mathit{zip}([a_0, \ldots, a_n], [b_0, \ldots, b_n]) = [(a_0, b_0), \ldots, (a_n, b_n)]$.

[19] For reading long concatenations of multiple functions, proceed from right to left and trace the types of the intermediate computations.

is required to compute an outcome of type Val_c.[20]
Now, two probability distributions of trajectories $pxs, pxs' : Prob(List(X))$ can e.g. be compared by

$$(compare \circ map^{\blacktriangle}(measureVal_o \circ map^{Prob}(aggregateVal_o \circ map^{List}(observe))))(pxs, pxs')$$

or

$$measureVal_c \circ map^{Prob}(compare) \circ map^{\blacktriangle}(map^{Prob}(aggregateVal_o \circ map^{List}(observe)))(pxs, pxs')$$

Again, it is not clear that these computations will return the same result.

It would also be interesting to consider non-deterministic variants of the second experiment. Moreover, one might want to compute not only one optimal solution but several that are "good enough" according to some metric and compare the corresponding choices for $S_\bullet$.

IPCC estimates. Using such and other methods, the IPCC gives estimates of value ranges for $\Delta T_{2\times CO_2}$ based on "Understanding of climate processes, the instrumental record, paleoclimates and model-based emergent constraints". According to the AR6 WGI report, it is considered as *very likely* [21] that the value of $\Delta T_{2\times CO_2}$ lies between 2°C and 5°C with *high confidence* in the lower and *medium confidence* in the upper bound As best estimate the report now gives a value of 3°C with $[2.5°C, 4°C]$ as *high confidence* range, while in AR5 a best estimate was not given and the *likely* range was stated as $[1.5°C, 4.5°C]$. (However, it is also said that the CMIP6 models exhibit "a larger range of climate sensitivity" and a higher average than the CMIP5 models and the AR6 best estimate. This behaviour is attributed to an amplifying cloud feedback.) Besides giving these ranges, the report also states that since AR5 "independent lines of evidence, including proxy records from past warm periods and glacial-interglacial cycles, indicate that sensitivity to forcing increases as temperature increases (TS.3.2.2)". The latter seems to indicate that "sensitivity to forcing" is to be understood not with respect to pre-industrial as in the IPCC's ECS definition, but in a more general sense in which reference states with different values of GMST can be considered.

4.2 Selected references

Origins and early studies:

- A very early study of the influence of the CO_2 content of the atmosphere on the mean surface temperature: Arrhenius [Arr96].
- Budyko [Bud69] and Sellers [Sel69] study EBM models of the radiative fluxes at the top of the atmosphere.
- The original definition of ECS stems from the *Charney report* [Cha+79] which also gave the first estimate range based on the models by Manabe/Wetherald [MW75] and Sellers[Sel69].
- Hansen et al. [Han+84] study climate sensitivity in three ways that have become standard "lines of evidence": by model simulation, from paleo data and from the instrumental record.
- An early review of climate sensitivity studies: Schlesinger [Sch83].

There is a huge body of literature with computations of forms of ECS and TCR based on different lines of evidence. In the last decade there have been efforts to systematise these results and to develop the notion of "climate sensitivity" to be more specific about underlying assumptions of individual computations of climate sensitivity. As studies have also shown a likely state dependence of ECS (e.g. [AGW15]), recent publications suggest extensions to the basic concept. The latter is

▶ **TiPES** Motivates Objectives 4.1 and 4.2

[20] Note that $aggregateVal_o$ and $aggregateVal_c$ might be seen as measures for the list monad.

[21] The IPCC uses a "calibrated language" for a consistent treatment of uncertainty estimates where natural language expressions like *likely*, *very likely*, *high confidence* are chosen according to guidelines described in [Mas+10].

the focus of TiPES WP4 "From Climate Sensitivity to a general theory of Climate Response across scales".
The following are papers that systematise and improve the study of climate sensitivity and more generally the climate response to external forcing:

- The PALEOSENS group [Roh+12] is concerned with a systematic treatment of ECS studies based on paleo data. To generalise the standard 2×CO_2 ECS notion they propose to use a climate sensitivity parameter per unit radiative forcing instead (the parameter $S_\bullet$ discussed above) and to make explicit the time-scale of the climate feedbacks that are considered in the sensitivity computation. Beyond the "Charney" definition of ECS which includes "fast" feedbacks, they propose ESS as a variant of ECS with the same general structure, but which also includes "slow" feedbacks.
- The authors of [Hey+16] study and discuss the problems with ECS in presence of tipping points.
- Knutti et al. discuss the shortcomings of linearity assumptions (as e.g. in Eq. 25) that are common in climate sensitivity studies in [KR15] and alternative metrics "beyond ECS" in [KRH17].
- A long recent report [She+20] on the assessment of "Earth's climate sensitivity" combines evidence from the understanding of feedback processes and both the historical and the paleo record with expert judgement. To transparently deal with uncertainty, the *storyline* approach [She+18] is employed, as already done in an earlier paper by some of the authors [Ste+16]

A concluding quote on the difficulty of "quantifying ECS" from [She+20]:

> Quantifying ECS is challenging because the available evidence consists of diverse strands, none of which is conclusive by itself. This requires that the strands be combined in some way. Yet, because the underlying science spans many disciplines within the Earth Sciences, individual scientists generally only fully understand one or a few of the strands. Moreover, the interpretation of each strand requires structural assumptions that cannot be proven, and sometimes ECS measures have been estimated from each strand that are not fully equivalent. This complexity and uncertainty thwarts rigorous, definitive calculations and gives expert judgement and assumptions a potentially large role.

5 Commitment

In the context of climate change, one comes across three different notions of *commitment*:

1. *climate change commitment* as a technical notion, used to quantify delayed responses of the climate system due to some form of inertia, e.g. warming that would result from past emissions even if CO_2 emissions were stopped immediately
2. *commitment* in the sense of contractual obligation as used in e.g. in the Kyoto protocol
3. *commitment* as a modal operator capturing a form of personal dedication to the fulfilment of a task

Our main focus here is on the first notion.

5.1 Climate change commitment

The first usage of this notion according to the AR4 WGI report is in [Ram88].

> "The inferred trace gas increases from the preindustrial era to the present have *committed* the planet to an equilibrium surface warming of about 0.6 to 2.4 K. Furthermore, at the current rate of increase in the trace gases, the *committed equilibrium warming* of the globe increases by about 0.13 to 0.5 K per decade." (*emphasis added*)

We see that Ramanathan's notion of commitment is defined in terms of an *equilibrium warming*. Wetherald et al.[WSD01] similarly consider a notion of *warming commitment*:

> "The difference between the realized warming at a given time and the warming of climate that would occur if the climate had an infinitely long time to adjust to that radiative forcing (i.e. the gap between the equilibrium and the realized temperature change for a given forcing) is referred to here as "the **warming commitment**".
>
> We re-examine the present day **warming commitment**, its future changes and other **committed climate responses**. [...] allows the **climatic commitment** of a wide range of variables to be examined [...]" *(emphasis as in the source)*

Wigley's 2005 paper "The climate change commitment" [Wig05] (according to the IPCC AR5 WGI contribution) is the source of the terminology and most of the variants of the *climate change commitment* notion in the IPCC glossary. On alternative terminology for commitment in earlier work Wigley writes:

> "For global-mean temperature, this is referred to as the unrealized warming [Han+85], residual warming [Wig84], or committed warming [WSD01]. Here, I use the term **warming commitment** or, to include sea level rise [WR93; SM99], **climate change commitment**." *(citations adapted)*

and on *equilibrium warming commitment*:

> "The **usual (or equilibrium) CC warming commitment at time t** is **the difference between the equilibrium warming for forcing at this time (ΔT_e) and the corresponding realized warming (ΔT_r),** $\Delta T_e - \Delta T_r$. This is related to the radiation-imbalance concept [Han+02; Pie03]. If ΔQ is the forcing to date, and if ΔQ_r is the forcing that gives an equilibrium warming of ΔT_r, then the radiation imbalance is $\Delta Q - \Delta Q_r$ ($\Delta Q - \Delta Q_r$ is approximately equal to the flux of heat into the ocean [Pie03]). Hence
>
> $$\Delta T_e - \Delta T_r = (\Delta Q - \Delta Q_r)(\Delta T2\times/\Delta Q2\times)$$
>
> where $\Delta Q2\times$ is the radiative forcing for a CO_2 doubling (about $3.7\mathrm{W\,m^{-2}}$) and $\Delta T2\times$ is the corresponding equilibrium global-mean warming. A central estimate of ΔQ (accounting for both natural and anthropogenic forcings) is about $1.7\mathrm{W\,m^{-2}}$, whereas ΔT_r is about 0.7°C. Given $\Delta T2\times = 2.6°\mathrm{C}$ [WR01], a central value for the current **equilibrium warming commitment** is about 0.5°C, with a corresponding radiation-imbalance estimate of $0.7\mathrm{W\,m^{-2}}$. These results are in accord with other estimates in the literature, but uncertainties are large." *(citations adapted)*

We see that the computation of equilibrium climate commitment presented by Wigley depends on the value of ECS. He motivates his above computation of equilibrium commitment by referring to "the radiation-imbalance concept". To our understanding, he assumes the relation between change in radiative forcing and change in temperature discussed in the beginning of Section 4, such that $(\Delta T2\times/\Delta Q2\times) = 1/S_\bullet$. The estimate $\Delta Q2\times \approx 3.7\mathrm{W\,m^{-2}}$ coincides with the estimate resulting from the Myhre et al. formula of Eq. (24).

Wigley posits the two scenarios that are still standard:

> "The assumption of constant atmospheric composition on which the warming commitment idea is based is clearly unrealistic, even as an extreme case of what might happen in the future. An alternative indicator of the **commitment to climate change** is to assume that the emissions (rather than concentrations) of radiatively important species will remain constant. This Report investigates the **constant-composition (CC) warming and sea level commitments**, the **constant-emissions (CE) commitments**, and the uncertainties in each."

He proposes to study time-dependent variants of commitment instead of just the equilibrium/asymptotic commitment [22]

> " Because it would take an infinite time for the unrealized warming to appear, a more useful definition makes the unrealized warming a time-dependent quantity, namely, the evolving changes in global-mean temperature that would result if atmospheric composition were kept constant at its present state [Wig84]. This is the definition I use here. Temperatures under this new definition tend asymptotically to the previous equilibrium commitment definition. The new definition can be applied equally to the **CC and CE commitments** and can be used for both temperature and sea level." *(citation adapted)*

Further geophysical variants. Wigley's definitions of commitment are based on a difference, while the AR5 WGI report [Kri+13] also mentions the usage of a ratio-based constant composition commitment, citing [Sto04; Mee+07; Sol+09; Eby+09]:

> "A measure of constant composition commitment is the fraction of realized warming which can be estimated as the ratio of the warming at a given time to the long-term equilibrium warming [...]"

Furthermore, the AR5 WGI report discusses the following form of commitment studied in [CC10] (and similarly in [Hel+10]):

> "Another form of commitment refers to climate change associated with heat and carbon that has gone into the land surface and oceans. This would be relevant to the consequences of a one-time removal of all of the excess CO_2 in the atmosphere and is computed by taking a transient simulation and instantaneously setting atmospheric CO_2 concentrations to initial (pre-industrial) values [...]"

Non-geophysical variants. There are also notions of commitment that are related to inertia in technological, societal and economical aspects. An overview is given in the AR5 WGI report [Kri+13]:

> "A **more general form of commitment** is the question of how much warming we are committed to as a result of inertia and hence commitments related to the time scales for energy system transitions and other societal, economic and technological aspects (Grubb, 1997; Washington et al., 2009; Davis et al., 2010). For example, Davis et al. (2010) estimated **climate commitment** of 1.3°C (range 1.1°C to 1.4°C, relative to pre-industrial) **from existing CO_2-emitting devices** under specific assumptions regarding their lifetimes. These forms of commitment, however, are strongly based on political, economic and social assumptions that are outside the domain of IPCC WGI and are not further considered here."

5.2 Other usages

Commitment also denotes an obligation following from entering a contract. E.g. in the following paragraph from the IPCC AR5:

> "[...] Parties with quantified emission limitations (and reduction obligations) in aggregate may have bettered their collective emission reduction target in the first *commitment* period, but some emissions reductions that would have occurred even in its absence were also counted. The Protocol's Clean Development Mechanism (CDM) created a market for emissions offsets from developing countries, the purpose being two-fold: to help Annex I countries fulfill their *commitments* [...]" *(emphasis added)*

[22] Apparently he has already used this approach in [Wig84] which he cites; this paper however seems to be unavailable online.

The above refers to the Kyoto protocol which has two *commitment periods*. The European Commission Website writes about the first commitment period:

> "In the first period of the Protocol (2008-12), participating countries *committed to* reduce their emissions by an average of 5% below 1990 levels." *(emphasis added)*

Of the non-technical meanings of *commit* and *commitment* listed in the *Oxford English Dictionary* item 6 seems to be closest to the usage in climate science:

> "6. a. The action or an act of obligating or binding oneself or another to a particular course of action, policy, etc.; the action of giving an undertaking, either explicitly or by implication. Also: an undertaking or pledge of this kind.
>
> [...]
>
> b. An act or course of action to which a person is bound or obligated; an obligation, responsibility; a liability; an engagement."

There is also recent game-theoretic work concerning a similar notion of *conditional commitment* [Hei19].

Deontic logic. Feltus and Petit [FP09] propose a formalisation of a commitment modality in standard deontic logic (a *modal logic* concerned with *obligation*, *permission* and related concepts). Their understanding of commitment is described as follows:

> "It appears that this option proposition, even if optional to an organizational accountability, could remain engaged toward a moral obligation that we call **Commitment**. This commitment could be defined as the act of binding itself (intellectually or emotionally) to a course of actions."

This seems more compatible with item 7 of the OED:

> "7. a. The state or quality of being dedicated to a cause, ideology, activity, etc.; the action of devoting oneself to something or someone; devotion, dedication."

5.3 Instances of climate change commitment

As we have seen, there are different instances of *climate change commitment*, depending on which forcing scenario is used, which feature of the climate system is of interest and whether an equilibrium or a transient version of the notion is considered.

Climate Change Commitment: Instances

- with respect to different features of a climate system:
  - temperature
  - in the hydrological cycle
  - extreme weather events
  - extreme climate events
  - sea level change
- with respect to specific scenarios:
  - constant composition
  - constant emission
  - zero emission
  - feasible scenario
- as transient or equilibrium notion

5.4 Computational structure of climate change commitment

Climate change commitment can be estimated similarly to climate sensitivity by model simulation, from:

- A model m with state space X and parameters $(f, p') : (\mathcal{T} \to P) \times P'$ where the first component denotes a forcing
- An initial time $t_0 : \mathcal{T}$ and state $x_0 : X$
- A numerical method for computing approximations $\mathit{trajectory}_m(n, x_0) = [x_0, \ldots x_n]$ for a time discretisation $[t_0, \ldots, t_n]$
- a reference time $t_{\text{ref}} : \mathcal{T}$ and state $x_{\text{ref}} : X$ (in which the climate system is not necessarily in radiative equilibrium)
- a forcing $f : \mathcal{T} \to \mathbb{R}$.

  The main variants of climate change commitment assessments listed in the glossary of the IPCC's SR15 [IPC18] differ only in the considered forcings: *constant composition*, *constant emissions*, *zero emissions* and *feasible* scenario, defined as functions

  $$\begin{aligned} &\mathit{constComposition} : \mathcal{T} \to \mathbb{R} \\ &\mathit{constEmissions} \quad : \mathcal{T} \to \mathbb{R} \\ &\mathit{zeroEmissions} \quad\; : \mathcal{T} \to \mathbb{R} \\ &\mathit{feasible} \qquad\quad\;\; : \mathcal{T} \to \mathbb{R} \end{aligned}$$

  which force the components of the system's state that represent the atmospheric concentration of CO_2 to

  - remain constant
  - change corresponding to a constant amount of anthropogenic emissions
  - change according to zero anthropogenic emissions
  - change according to the minimal amount of anthropogenic emissions that is judged as feasible

  If another forcing scenario f is used, the idea that commitment quantifies the response of the climate system when an anthropogenic forcing is stopped or at least does not increase anymore suggests that f should be a non-increasing function:

  $$\forall t, t' : \mathcal{T}, t < t' \Rightarrow f(t) \geqslant f(t')$$

- a function $\mathit{observe} : X \to \mathit{Val}_o$ observing the feature of a climate state we are interested in, e.g. again a function that computes the GMST as in Section 4 or a measure of the sea level

  $$\mathit{gmst} : X \to \mathbb{R}$$

  $$\mathit{sealevel} : X \to \mathbb{R}$$

- a function that aggregates a trajectory to a representative state (as in Section 4)

  $$\mathit{aggregateX} : \mathit{List}(X) \to X$$

- and a comparison function $\mathit{compareVal}_o : \mathit{Val}_o^2 \to \mathit{Val}_c$ that allows us to compare the resulting values of the feature of interest. Typically Val_o and Val_c are numerical types and the comparison function computes the difference of two numbers or their ratio.

The computation then follows the structure of a comparison experiment. To our understanding the difference to the CS experiments described in Section 4 is that here the reference state which is used for comparison is not supposed to be in equilibrium.

We have seen above that Wigley distinguishes between transient and equilibrium commitment experiments. From a theoretical perspective, when integrating a dynamical system indefinitely "until an equilibrium is reached", it is not guaranteed that it will ever do so and the computation terminate. In practice however, the computations performed for an equilibrium experiment will rather correspond to a transient experiment for a number of time steps in which the processes represented in the model are expected to equilibrate (or otherwise the equilibrium value may be estimated as for the EECS notion):**TODO** [23]

- **transient:** To obtain a quantity of transient commitment $ccc_{\text{tr}} : \mathit{Val}_c$ for a reference state x_{ref} at time t_{ref} that is reached after n_{ref} times steps with respect to a particular time t, which is reached after n_{tr} time steps in the numerical simulation, compute the trajectories

  $$xs_{\text{tr}} = \mathit{trajectory}(n_{\text{tr}}, x_0) \qquad \text{and} \qquad xs_{\text{ref}} = \mathit{trajectory}(n_{\text{ref}}, x_0)$$

  and then

  $$ccc_{\text{tr}} = (\mathit{compareVal}_o \circ \mathit{map}(\mathit{observe} \circ \mathit{aggregateState})(xs_{\text{tr}}, xs_{\text{ref}}).$$

- **equilibrium:** To obtain a quantity of equilibrium commitment $ccc_{\text{eq}} : \mathit{Val}_c$, we need either to choose a number n_{eq} of computation steps that is sufficiently large to let the model reach an equilibrium or a function $\mathit{estimate}_{\text{eq}} : \mathit{List}(X) \to X$ to compute an equilibrium state (as in Section 4). In the first case, the computation does not differ from the transient case. In the second case, one computes an equilibrium notion of commitment by first computing the trajectory

  $$xs_{\text{ref}} = \mathit{trajectory}(n_{\text{ref}}, x_0)$$

  and then

  $$ccc_{\text{eq}} = (\mathit{compareVal}_o \circ \mathit{map}(\mathit{observe}))(\mathit{estimate}_{\text{eq}}(xs), \mathit{aggregateX}(xs_{\text{ref}})).$$

The commitment experiment suggests to slightly update the comparison experiment as defined in Eq. (29) to allow the two trajectories that are to be compared to have different lengths:

$$\begin{aligned} &\mathit{comparisonExperiment} : (\mathit{List}^2(X) \to \mathit{Val}_c) \times \mathbb{N}^2_{<N} \times X^2 \to \mathit{Val}_c \\ &\mathit{comparisonExperiment}\,(\mathit{compare}, n, n', x, x') \\ &\qquad = \quad (\mathit{compare} \circ \mathit{map}(\mathit{trajectory}))((n, x), (n', x')) \end{aligned} \tag{32}$$

Then a commitment experiment can be expressed as

$$\begin{aligned} &\mathit{commitmentExperiment} : \mathbb{N}^2_{<N} \times X^2 \to \mathit{Val}_c \\ &\mathit{commitmentExperiment}\,(n, n', x, x') \\ &\qquad = \quad \mathit{comparisonExperiment}(\mathit{compareVal}_o \circ \mathit{map}(\mathit{observe} \circ \mathit{aggregateX}), n, n', x, x') \end{aligned} \tag{33}$$

[23] Michel: Maybe rearrange a little?

Remarks. As for the CS experiments, we could consider non-deterministic variants and different algorithmic ways of performing measuring, aggregation and comparison operations.
The notion of commitment seeks to quantify how much more climate change is inevitably to be expected at some moment in time – however it crucially depends on the future forcing that is considered. This suggests to consider a notion of multi-scenario commitment in which the considered forcing scenarios represent different policy options, possibly weighted by a judgement about feasibility. With the expected presence of tipping points in the climate system (see Section 6), one might also ask under which forcing scenarios present day or future commitment for some climatic variable entails the crossing of such tipping points. These ideas are explored in TiPES Deliverable 6.3. In [Bot+21] (part of TiPES D6.2), the idea of commitment is used in a stylised way as criterion for partitioning the state space of a dynamical system into desirable and undesirable regions, reminiscent of Heitzig et al.'s topological classification in [Hei+16] and the planetary boundaries/safe operating space concept of [Roc+09].

5.5 Selected references

- Ramanathan [Ram88] is the first to use the term "committed" in the sense of *climate change commitment*.
- Whetherald et al. [WSD01] consider a notion of *warming commitment*.
- Wigley's paper [Wig05] is the source of the terminology and most of the variants of the *climate change commitment* notion in the IPCC glossary.
- The glossary of the IPCC report Global Warming of 1.5°C[IPC18] includes an entry on *Climate change commitment*, distinguishing between different instances depending on the forcing scenario considered.
- Chapter 12 of IPCC AR5[Kri+13] discusses *climate change commitment*.
- Plattner et al. [Pla+08] use different EMICs to compute the kinds of climate change commitment discussed by the IPCC.
- Lenton et al. [Len+08; Len+19] use the word "committed" in the context of tipping points and tipping elements (cf. Section 6.4) but no explanation is provided to what they explicitly mean by committed.
- Pattyn et al. [Pat+18, Box 2] relate "climate commitment" and tipping points in a study of Greenland and Antarctic ice sheets at 1.5C warming, without further definition of the notion, but likely referring to *sea level rise commitment* in the sense of the IPCC.
- Heitzig [Hei19] uses the notion of *conditional commitment* in the sense of contractual obligations in a game-theoretic setting.
- Feltus and Petit [FP09] propose a *commitment modality* for deontic logic.

6 Abrupt Change, Tipping Point, Tipping Element

This section is concerned with notions related to *abrupt changes* in the Earth system, in particular *tipping point* and *tipping element*.

Definitions. The IPCC glossary entries provide the following definitions (cf. Appendix I.6):

- an *abrupt change* in a system is a change that "takes place substantially faster than the rate of change in the recent history of the affected component of a system"
- if the "abrupt change occurs because the system state actually becomes unstable, such that the subsequent rate of change is independent of the forcing" it is referred to as *tipping point*

- A *tipping point* is “defined as a critical threshold beyond which a system reorganizes, often abruptly and/or irreversibly”
- where a “perturbed state of a dynamical system is defined as *irreversible* on a given timescale, if the recovery from this state due to natural processes takes substantially longer than the timescale of interest” *(emphasis added)*

This description of *abrupt change* corresponds to the definition given in a 2002 *National Research Council (NRC)* report on “abrupt climate change” [BC+02]:

Abrupt climate change

Definition from [BC+02, p.14]:

> “Technically, an abrupt climate change occurs when the climate system is forced to cross some *threshold*, triggering a *transition* to a new state at a rate determined by the climate system itself and faster than the cause.” *(emphasis added)*

It should be noted that the usage of “state” in the above descriptions apparently differs from how we have used it in the previous sections. Here “state” rather seems to refer to a *region* of the system’s state space which contains states that are qualitatively distinct from the states in other regions. For a system with state space X in our sense, such a region $R \subset X$ could be described with a predicate $P : X \to \mathit{Prop}$ such that for all states $x : X$,

$$x \in R \Leftrightarrow P(x).$$

A possible terminology for such qualitatively different regions of the state space could be to consider them as *macro states*, while the individual elements of the state space of a system could be referred to as *micro states*.

Example 6. Alley et al. [All+03] illustrate the idea of abrupt change by the concrete example of flipping a canoe, while introducing the notions of *trigger, amplifier, globaliser* and *source of persistence* as typical ingredients that bring about such changes after crossing a *threshold*:

> “Systems exhibiting *threshold* behavior are familiar. For example, leaning slightly over the side of a canoe will cause only a small tilt, but leaning slightly more may roll you and the craft into the lake. An abrupt change, of a canoe or the climate, requires a *trigger*, such as you leaning out of a canoe; an *amplifier* and *globalizer*, such as the friction between you and the canoe that causes the boat to flip with you; and a *source of persistence*, such as the resistance of the upside-down canoe to being flipped back over.” *(emphasis added)*

Alley et al. also point out that

> “Such large and rapid threshold transitions between distinct states are exhibited by many climate models, including simplified models of the oceanic thermohaline circulation [Sto61], atmospheric energy-balance models [Sel69], and atmospheric dynamical models exhibiting spontaneous regime changes [Lor63].” *(citations adapted)*

Kuehn [Kue11] formalises a notion of *critical transition* using the theory of slow-fast dynamical systems. He lists the following attributes of critical transitions (citing [Sch+09]):

- occurrence of an abrupt qualitative change in the system
- the change occurs rapidly relative to the regular system dynamics
- the system crosses a special threshold near a transition
- the new state of the system after the transition is “far away” from its previous state.

Ashwin et al.[Ash+12] point out that the more recently popularised notion of *tipping point* is related to a much older question in climate science:

> "The recent interest in tipping points is related to a long-standing question in climate science: to understand whether climate fluctuations and transitions between different 'states' are due to external causes (such as variations in the insolation or orbital parameters of the Earth) or to internal mechanisms (such as the oceanic and atmospheric feedbacks acting on different time scales)."

In the following, we will briefly address abrupt changes in paleo data, stability and tipping in climate models and the study of different kinds of tipping behaviour in deterministic and stochastic dynamical systems.

6.1 Abrupt Changes in Time Series

The probably simplest example of a "regime change" is the transition between "warm" and "cold" states of the Earth between *greenhouse periods* (in which no continental glaciers exist) and *glaciations* (ice ages, "Snowball Earth") with an $\approx 100,000$ yr periodicity documented in proxy records [Sha00]. Very prominently, the notion of *abrupt change* in time series of paleo records is used for events that occurred during glacial or interglacial periods (colder or warmer periods within a glaciation). Alley et al. [All+03] describe such changes as follows:

> "For example, global-mean temperature changes of perhaps 5°C to 6°C over ice-age cycles [Bro02] are generally believed to have resulted from small, globally averaged net forcing [All+03, elaboration (5)]. More surprisingly, regional changes over 10 years without major external forcing were in many cases one-third to one-half as large as changes over the 100,000-year ice-age cycles [Bro02; Sto00]." *(citations adapted)*

and

> "Regional climate changes of as much as 8° to 16°C [Sto00; Sev+98] occurred repeatedly in as little as a decade or less" *(citations adapted)*

Example 7. *Dansgaard-Oeschger (D-O) events.* The first concrete evidence that the climate system has undergone abrupt changes in the past was found in Greenland ice core records [Dan+82; Dan+84; Oes+84; Dan+93; Ank+93] in the 1980s. Stocker [Sto00] writes:

> "The most detailed and continuous information about climate variability comes from the Greenland ice cores (Fig. 4). Hans Oeschger [Oes+84], Willy Dansgaard [Dan+84] and colleagues were the first to recognise the climatic significance of short interstadials (warming events) during the last glacial period; they were numbered consecutively [Dan+93] and later named 'Dansgaard/Oeschger Events [BD89] (see Fig. 5). All events exhibit a striking similarity in their temporal evolution: cooling extends generally over many centuries to about 3 kyr, while warming is abrupt and occurs within years or decades. This suggests that one common mechanism may be responsible for these climate swings. The recurrence time for the shorter D/O events is of the order of 1000 years."
> *(citations adapted, figure references for the original paper)*

Similar recurring oscillations have been found in other records:

- *Heinrich events* are recognised as sequences of iceberg discharges that left [ice rafted?] [debris?] in the North Atlantic [Hei88; Bro+92; Bon+93]/ They are linked to Antarctic Warming Events [WFR09] and thought to be linked to the D-O events

- *Bond events* are climate fluctuations in Holocene records that have a much smaller amplitude than D-O events [Bon+97]

Section 4.1 of [Sto00] provides an overview of "abrupt climate change documented" in paleo records. These proxy records are considered as evidence that there were "repeated, large, abrupt shifts in Northern Hemisphere climate during the last ice age" [All00; Bar+11]. The implications of such abrupt shifts for present day climate change started (at the latest) to be discussed around 2000 with a National Research Council report and several accompanying papers [All00; All+03; BC+02].

There have been diverse efforts to explain the causes of D-O events. The question concerning internal vs external causes as stated in the Ashwin et al. quote in the introduction of this section is mirrored in these efforts. To this day, there is no consensus as to the cause of the D-O events, but it is commonly accepted that they are linked to changes in the location and intensity of water convection in the North Atlantic, with global impacts both on the atmospheric and the oceanic circulation measured, among others, in the isotopic signatures of the south Asian monsoon and Antarctic precipitation.

Methods. Different approaches to paleoclimate modelling in general are described in [Cru12]. A method to explore the possible underlying dynamics of D-O events is described and employed by Lohmann and Ditlevsen [LD19]. Ditlevsen et al. [DAS07] explicitly discuss how DO events and periodicity have been defined in the literature.

6.2 Multi-stability and tipping in conceptual climate models

It was known long before the more recent discussion of *tipping points* that "large and rapid threshold transitions between distinct states are exhibited by many climate models" [All+03].

Early examples of such models include those by Stommel [Sto61][24], Budyko[Bud69], Sellers[Sel69] and Lorenz [Lor63] mentioned in Section 2. These are conceptual models that are studied using dynamical systems and bifurcation theory [Kuz13; Arn+13]. A model is called *multi-stable* if it has more than one stable equilibrium.

Stommel's model concerns the strength of the thermohaline circulation in the Atlantic and the qualitative distinction between possible equilibria is whether the circulation is "on" or "off".[25]

The Budyko and Sellers model are EBMs modelling the Earth's radiative fluxes at the top of the atmosphere. An early stability study of such a model was conducted by Ghil [Ghi76]. Another simple EBM in which the number of equilibrium states depends on the value of a certain parameter was proposed by Fraedrich[Fra79] and extended to a stochastic model by Sutera[Sut81].

Prototypically, such "large and rapid threshold transitions" can be explained by the presence of *bifurcations* in the model: a bifurcation occurs "when a small smooth change to a parameter (or parameters) of a system causes a sudden qualitative or topological change to its behaviour" [Len13], e.g. the set of equilibrium solutions of the system changes. However, not all forms of *tipping point behaviour* are caused by a bifurcation. Lenton writes [Len13]:

> "There are [...] several mathematically distinct potential sources of *tipping point behavior* (where a small change gives rise to a large response in a system)." *(emphasis added)*

and defines tipping point as follows:

Tipping Point

Definition following [Len13]:

> A tipping point is a point at which a small perturbation can cause a qualitative change in the future state of a system.

The formulation "the future state" seems again to refer to a region of the state space.

[24] The variant of this model discussed in Examples 1 and 4 exhibits the same multi-stability.

[25] According to the appendix of [Mar00], Stommel's 1961 model "went virtually unnoticed for 25 years" while "in 1982 another box model was independently proposed" (the Rooth model [Roo82]) "that explained how a two-hemispheric THC [...] might become unstable". Finally, it "was extensively applied to the steady-state pole-to-pole circulation" by Marotzke [MW91; Mar94] and Rahmstorf [Rah96].

Different causes of tipping. In the literature, different kinds of tipping point behaviour are considered which allow to classify critical transitions by their cause. To our knowledge, the following have been discussed:

Classification of Tipping

- *B-tipping*: bifurcation-induced [Ash+12]
- *N-tipping*: noise-induced [Ash+12]
- *R-tipping*: rate-dependent [Ash+12] (e.g. the "compost-bomb instability" [Wie+11])
- *S-tipping*: shock-induced [HF20]
- *P-tipping*: phase-sensitive [ATW21]

B-tipping may occur in a dynamical system m with parameters of type P for which a critical value $p_{\mathrm{cr}} : P$ exists at which the number or stability of equilibrium solutions of the system changes.
When the value of this parameter changes due to a forcing $f : \mathcal{T} \to P$ [26] such that the critical value is crossed, the trajectories of the forced system may pass through states that approach different equilibrium solutions before and after the crossing of the critical value. These different equilibrium solution may in turn correspond to very different observations.

N-tipping may occur in stochastic dynamical systems with multiple equilibrium solutions when noise can cause a trajectory to transition between two states which are attracted by different equilibrium solutions.

R-tipping may occur if such a transition between states leading to different equilibria is triggered when the forcing exceeds a critical rate of change.

S-tipping may occur if such a transition between states leading to different equilibria can be caused by a large perturbation (a shock or extreme event) in the forcing.

Finally, *P-tipping* may occur in cyclic systems in which the possibility of transitioning to a qualitatively different state space region may depend not only on the crossing of a critical parameter value but also on the current phase of the cycle.

Ashwin et al. [Ash+12] show that dynamical systems can exhibit B-, N- and R-tipping behaviour independently or in combination. A simple EBM, the Faedrich-Sutera model [Fra79; Sut81], is shown by Ashwin et al. to exhibit B-, N- and R-tipping.

6.3 Tipping in complex climate models

With the increasing evidence that thresholds for climate tipping points could be crossed as consequence of anthropogenic forcing, concerns have arisen that state of the art climate models may be "too stable": if they are unable to adequately capture abrupt changes as observed in paleo data, they might not reliably predict abrupt changes in the near future either [Val11]. This concern links to the key question stated in the TiPES proposal for WP2: "Are our models too stable?".
In a recent review article Brovkin et al. [Bro+21] state that

> "For Earth system modellers, the main task is the further improvement of their models and coupled atmosphere-ocean-biosphere-cryosphere processes. Good progress is being made with Earth System Models[Fla11]; they are capable of simulating some abrupt changes, especially in the cryosphere, during the past century and in future projections[Dri+15]. However, they are challenged by attempts to reconstruct abrupt events

[26] and at each time $t : \mathcal{T}$ the trajectory of the system is thus determined by $m_{f(t)}$

> that are well documented from the past, including meltwater pulses due to ice sheet collapses[BHD10], the rapid release of CO_2 during deglaciation[Mar+14] and abrupt climate and vegetation changes in North Africa during the termination of the AHP[Dem+00]."
> *(citations adapted from the paper)*

This provides motivation for the objectives of TiPES WP2: Assess stability for various ESMs and EMICs (see box on climate models) and facilitate evaluation of model behaviour relative to proxy data prepared by WP1.

▶ **TiPES** Links to Objectives 2.1, (2.2), 2.3

6.4 Tipping Points and Tipping Elements

While the notion of *tipping point (TP)* [27] in the sense of *abrupt change* already appeared earlier, the notion of *tipping element (TE) in the Earth's climate system* was introduced in Lenton et al.'s paper [Len+08] (following Lenton and Schellnhuber's commentary [LS07]). Roughly, a tipping element is a subsystem of the Earth system that "may pass" a tipping point. Being a TE is a mathematical property of a dynamical system.

However, Lenton et al. restrict their focus to a class of such subsystems which they judge as *policy-relevant*. Thus, the definition of policy-relevant TEs consists of two parts: one part defining mathematical properties that a TE must possess, and one part with subjective conditions. The latter might be chosen according to what a policy-maker deems relevant and might be seen as value judgements. The authors point out that in previous literature[28] , "abrupt changes" were described as occurring

> "when the climate system is forced to cross some threshold, triggering a transition to a new state at a rate determined by the climate system itself and faster than the cause"

and consider this a case of a bifurcation with an implicit focus on equilibrium properties and implying "some degree of irreversibility". They intend to give a definition that is broader than the above:

> "(i) we wish to include nonclimatic variables; (ii) there may be cases where the transition is slower than the anthropogenic forcing causing it; (iii) there may be no abruptness, but a slight change in control may have a qualitative impact in the future; and (iv) for several important phase changes, state-of-the-art models differ as to whether the transition is reversible or irreversible (in principle)."

The paper then gives a description of when a component of the Earth system is considered a tipping element exhibiting a tipping point. For a "full formal definition", the authors point to their supplementary material Appendix 1, "Formal Definition of a Tipping Element and its associated Tipping Point". We revisit the formal definition below. An informal short definition following the one given in the main paper is: [29]

Tipping Element

Let Σ be a subsystem of the Earth system associated with a specific region or collection of regions of the globe and at least subcontinental in scale (length scale of order $\approx 1000km$).

Σ is called a *tipping element* if

- the control parameters of the system can be combined into a single control of type P and
- there exists a critical control value $\rho_{\mathrm{cr}} : P$

[27] Lenton et al. do however not cite a scientific paper for this notion but the book by Gladwell [Gla00] that popularised the notion of "tipping point". Discussions and critical assessments concerning the usage of the "tipping point" notion in climate science can be found in [RN09; Rus11; Rus15; VHS18].

[28] citing [BC+02] as in the introduction of this section

[29] The notions "control parameter" and "control value" can simply be understood as "parameter" and "parameter value". We conjecture that they are used to indicate that the the parameter values in the systems they are interested in are influenced by policy decisions, and thus subject to being controlled.

such that any significant[a] deviation of the control value from ρ_{cr} leads to a qualitative change in the value of a crucial system feature after some observation time, measured wrt a reference value of the feature at the critical value.

[a] where a deviation is considered as “significant” if it is large relative to deviations caused by internal variability of the system

Interestingly, the definition does not explicitly use the notion *tipping point*. The informal introduction to the paper states

> The tipping point is the corresponding critical point – in forcing and a feature of the system – at which the future state of the system is qualitatively altered.

The supplementary material of [Len+08] states (on p.3, following the formal definition in which the notion *tipping point* does not occur)

> [...] our definition of a tipping element and its associated tipping point at the critical value

which seems to indicate that the critical value is considered as the tipping point. This understanding is confirmed in [Len12]:

> “In this definition [Len+08]. the critical threshold (ρ_{cr}) is the tipping point, beyond which a qualitative change occurs.”

In Table 1 of [Len+08], the different TEs are classified with respect to properties of the dynamical systems describing them:

Classification of Tipping Elements

A tipping element is classified by the

- feature of interest of the system with direction of change (increase or decrease)
- control parameter(s)
- critical value(s)
- transition timescale
- global mean warming at which the critical value might be reached
- key impacts

In this definition, the first four items are related to the mathematical definition of TEs, while the last two contain additional information that may help to judge the policy relevance of a TE.
Based on their criteria, Lenton et al. identified the following subsystems as instances of their definition of tipping elements in the Earth system [Len+08, Table 1]:[30]

▶ **TiPES**
Link to Objectives 1.3, 3.1, 3.2, 3.3, 3.5

Policy-relevant potential Tipping Elements: Lenton et al. Examples

- Arctic summer sea ice ♠
- Greenland ice sheet (*GIS*) ♠
- West Antarctic ice sheet (*WAIS*) ♠
- Atlantic thermohaline circulation (*THC*) ♠
- El Niño-Southern Oscillation (*ENSO*)

[30] (♠ marks TEs that are particularly in the focus of TiPES)

- Indian summer monsoon (*ISM*) ♠
- Sahara/Sahel and West African monsoon (*WAM*)
- Amazon rainforest ♠
- Boreal forest

Example 8. [Len+08, Table 1] characterises the THC as tipping element by the following data:

- feature of the system: amplitude of the overturning stream function, decreasing
- control parameter: freshwater input to the North Atlantic
- critical value: +0.1–0.5 Sv
- transition timescale: 100 yr (gradual)
- expected at global mean warming: +3–5 °C
- key impacts: regional cooling, sea level, ITCZ shift[31]

Lenton et al. also analysed some potential tipping elements which however did either not fulfil one of their TE criteria or there was not sufficient data at the time of writing and further research would be required. To our knowledge, the latest update of the assessment at the time of writing is presented in [McK+; Arm+].

6.5 Mathematical definition.

A mathematical definition of the notion *tipping element* is given in the supplementary material of [Len+08]. The formal framework in which the definition is stated is apparently considered as standard and not introduced in detail. The following is how we understand the essence of the definition translated to the notation of the current document:

Being a tipping element is a property of a subsystem Σ of the climate system. The following data is needed to represent Σ:

- a model of Σ with state space X such that forcing is represented by one time-dependent parameter of type $\mathcal{T} \to \mathbb{R}$ (thought of as *control path* for the system). [32]
- a start time (thought of as "now") $t_0 : \mathcal{T}$
- an initial state $x_0 : X$
- for each control path $p : \mathcal{T} \to \mathbb{R}$ the function $m_p : \mathcal{T} \to X$ defined by the model (with t_0 and x_0)

Then Σ is a *tipping element* iff there exist

- a system feature of interest given by a function [33]

$$F : X \to \mathbb{R}$$

- an "ethical" time span $T_E : \mathcal{T}$ $\qquad (T_E > 0)$

[31] The *Intertropical Convergence Zone* is the tropical rain belt where most of the rain on Earth falls [ABS16].

[32] The codomain type of the forcing, and similar for the system feature F below, is not explicitly stated by Lenton et al. Since it must be possible to add and subtract values of these types, to take their absolute value and to compare them, we used $\mathbb{R}$ as a reasonable choice. But the original definition might be intended to be more general such that any type with the necessary structure could be used.

[33] In [Len+08] this is said to be a projection from high-dimensional state space onto the component of interest. It is similar in spirit the function *observe* we used in the previous sections.

- a "critical" time $t_{\text{cr}} : \mathcal{T}$ $\quad (t_{\text{cr}} < t_0 + T_E)$
- a "critical" control path $p_{\text{cr}} : \mathcal{T} \to \mathbb{R}$
- an "exceedance time" $T_R : \mathcal{T}$ $\quad (T_R < t_0 + T_E - t_{\text{cr}})$
- a value $\delta : \mathbb{R}$ describing a small variation of the control value $\quad (\delta > 0)$
- a reference control path $p_{\text{ref}} : \mathcal{T} \to \mathbb{R}$
  such that $\quad \forall t \in [t_0, t_0 + T_E],\ p_{\text{ref}}(t) < p_{\text{cr}}(t_{\text{cr}})$
  and $\quad \forall t \in [t_{\text{cr}}, t_{\text{cr}} + T_R],\ |p_{\text{ref}}(t) - p_{\text{cr}}(t_{\text{cr}})| < \delta/2$
- a time scale T_v to filter variability $\quad (T_v \ll T_E)$
- a value $\hat{F} : \mathbb{R}$ $\quad \hat{F} > 0$
  that expresses a qualitative change in the value of the feature F

such that

I. Every control path which over the ethical horizon $[t_0, t_0 + T_E]$ is in the $\delta/2$ neighbourhood of p_{ref} and does not exceed $p_{\text{cr}}(t_{\text{cr}})$, compared to the reference path only leads to small changes (relative to $\hat{F}$) of the value of the feature F averaged over T_v-windows in the interval $[t_{\text{cr}}, t_0 + T_E]$:

$$(\forall t \in [t_0, t_0 + T_E], \quad |\, p(t) - p_{\text{ref}}(t)| < \delta/2 \quad \wedge \quad p(t) < p_{\text{cr}}(t_{\text{cr}}))$$
$$\Longrightarrow$$
$$\forall t \in [t_{\text{cr}}, t_0 + T_E], \quad |\langle F \rangle_t(m_p) - \langle F \rangle_t(m_{p_{\text{ref}}})| \ll \hat{F}$$

II. Any control path $p : \mathcal{T} \to \mathbb{R}$ that exceeds the critical value ρ_{cr} by δ at least in the interval $[t_{\text{cr}}, t_{\text{cr}} + T_R]$, leads within $[t_{\text{cr}}, t_0 + T_E]$ to the observation of a qualitative change $\geqslant \hat{F}$ relative to its development under the reference path p_{ref}:

$$(\forall t \in [t_{\text{cr}}, t_{\text{cr}} + T_R], \quad p(t) \geqslant p_{\text{cr}}(t_{\text{cr}}) + \delta)$$
$$\Longrightarrow$$
$$\exists T \in [t_{\text{cr}}, t_0 + T_E], \quad \langle F \rangle_T(m_p) - \langle F \rangle_T(m_{p_{\text{ref}}}) \geqslant \hat{F}$$

where $\langle F \rangle_t(x) : \mathbb{R}$ (with $x : \mathcal{T} \to X$) denotes the T_v-*moving average* of $F \circ x$ at time $t : \mathcal{T}$. [34]
Lenton et al. remark that this definition simplifies if T_R and T_E are long enough to let the system reach an equilibrium. They say that the system is now to be assumed as autonomous, which we understand in the sense that now the model is not parameterised by a time-dependent forcing but by a constant control parameter. Denoting the equilibrium value of the system feature F reached by the system with a control parameter $\rho : \mathbb{R}$ by $F_{eq}(m_{\text{const}(\rho)}) : \mathbb{R}$[35], they define:
Σ is a tipping element if there exists a control parameter with a critical value $\rho_{\text{cr}} : \mathbb{R}$ such that

$$|F_{\text{eq}}(m_{\text{const}(\rho_{\text{cr}}+\delta)}) - F_{\text{eq}}(m_{\text{const}\rho_{\text{cr}}})| \geqslant \hat{F}.$$

For reference, we also recall the definition given by Lenton et al. to define under which conditions a tipping element is considered as policy-relevant.
A tipping element Σ as defined above is called a *policy-relevant tipping element*, iff the following two conditions are fulfilled:

[34] The exact definition of $\langle F \rangle$ is not given in the paper, but we would expect

$$\langle F \rangle_t(x) = \frac{\displaystyle\int_{t-\frac{T_v}{2}}^{t+\frac{T_v}{2}} x(t)\, dt}{T_v}.$$

[35] where $\text{const}(\rho) : \mathcal{T} \to \mathbb{R}$ is the constant function that returns ρ for any input

III. the development of the control path within a political time horizon $T_P \approx 100\,\text{yrs}$ (driven by human interference) determines that a critical state will be reached at some point within T_E. "This may either be the case if the critical state is reached already within the political horizon, or if it would be reached at a later point in time in the absence of policies enacted during the political horizon to prevent the system from reaching its critical state."

IV. the qualitative change $\hat{F}$ of the value of the system feature F "could significantly affect human welfare on at least a sub-continental scale, or could compromise the overall mode of operation of the Earth system, or would entail the loss of a unique value of the biosphere".

Lenton et al. suggest the following concrete values or rules for the choice of T_v, T_E, δ and $\hat{F}$:

- $T_v \approx 10\,\text{yrs}$
  ("A reasonable choice for the time scale to filter variability in the system feature F [...] assuming that higher-frequency variations [...] are not relevant for the assessment whether or not a qualitative change $\hat{F}$ has occurred")
- $\delta \approx 0.2°\text{C}$ "for the particular case of annual global mean temperature"
- $T_E \approx 1000\,\text{yrs}$ "beyond which changes in the Earth system may not matter for current policy considerations"
- "$\hat{F}$ should be determined by considering associated impacts that fulfil requirement IV" or more general $\hat{F}$ should be "significantly larger than the standard deviation of natural variability" of the feature F on the time scale T_v

Remarks. While T_v, δ, T_E and $\hat{F}$ are existentially quantified in the definition of TE, the latter recommendations concerning the choice of concrete values for these variables suggest to restrict them to fixed values. To our understanding, the definition in this form has the advantage that it first describes the notion of TE in a purely mathematical sense which is applicable to a dynamical system independently of any physical and/or political interpretation. The policy-relevant part of the definition then restricts the class of TEs a posteriori. However, this approach makes the definition also rather difficult to parse. Maybe a simplified version of policy-relevant TE could instead be parameterised by F, T_E and $\hat{F}$ as political parameters and T_v, δ as physical parameters? Then a TE with respect to $(T_E, F, \hat{F}), (T_v, \delta)$ would be witnessed by choices of $t_{\text{cr}}, p_{\text{cr}}, T_R, p_{\text{ref}}$ that fulfil the coherence conditions and conditions I+II.

Another possibility to make the definition of TE more readable might be to introduce predicates like Near, Exceeds, FarAway etc. that describe the properties for control paths and trajectories used in conditions I+II.

6.6 Selected references

- Early multi-stable models: [Sto61], [Sel69], [Lor63], [Bud69]
- Early papers concerning D-O events: [Dan+82; Dan+84; Oes+84; Dan+93; Ank+93]
- Stocker [Sto00] gives an extensive overview of evidence for abrupt climate change in paleo records.
- Early papers addressing abrupt climate change more generally: [All00; All+03; BC+02]
- The original paper introducing the notion of *tipping element* is [Len+08]. A recent "updated assessment" of TEs is [Arm+].
- Different causes for tipping are introduced and studied in [Ash+12], [HF20], [ATW21].
- Kuehn [Kue11; Kue13] formalises the notion of *critical transition* in the framework of fast-slow systems.

- Valdes [Val11] discusses whether state of the art complex climate models are "too stable".
- The EU Horizon 2020 project COACCH has studied the potential impacts of climate and socio-economic tipping points for Europe [Gin+18; Trö+17].
- Recent work on "positive" non-climate tipping points: [Len20; Win+20; SL21]
- Russill et al. [RN09; Rus11; Rus15; VHS18] critically assess the usage of "tipping point" as *generating metaphor* in climate change communication.
- Lately there has been much interest in possible *tipping cascades*, e.g. [Krö+20; Loh+21; Klo+21; Bro+21; Wun+21].

7 Early Warning Signal

The observation that the climate system has undergone abrupt transitions in the past does not only trigger the question whether it might do so in the future, potentially as a consequence of human behaviour. It also suggests to ask whether there are ways to predict (and if possible prevent) imminent abrupt changes. The question of predictability has already been raised with the first papers explicitly focusing on abrupt climate change, e.g. in the context of the relation between D-O events and the thermohaline circulation: Marotzke's paper [Mar00] includes paragraphs "Would we Know Were It Happening Today?" and "A Different Brand of Predictability", and [Sto00] concludes with the importance of efforts to detect "early signs of sustained thermohaline circulation changes in the near future". With respect to ecosystems, the question of the predictability of approaching critical transitions has been posed even much earlier: Wissel [Wis84] discusses "a universal law for the characteristic return time near thresholds":

> "(a)[...] we search for a general property of ecosystems which is specific for the neighbourhood of thresholds; (b) [...] we ask if it is theoretically possible to use this property for the prediction of the position of a threshold"

In climate science, Tziperman [Tzi00] and Knutti/Stocker [KS02] describe the behaviour of the thermohaline circulation close to an instability threshold (the former based on simulations with a comprehensive climate model, the latter with a model of intermediate complexity). Consecutively, a variety of methods is proposed for the detection of approaching critical transitions. An overview of the proposed methods (illustrated by an application to simulated ecological data) is given in [Dak+12] (see also below). A more recent discussion on methods can be found in the introduction of [BBA20].
The notion *Early Warning Signal* seems to have been introduced by Dakos et al. in [Dak+08]:

> "our way to detect slowing down might be used as a universal early warning signal for upcoming catastrophic change"

and more prominently in Scheffer et al.'s "Early-warning signals for critical transitions" [Sch+09], in which the authors write

> "work in different scientific fields is now suggesting the existence of generic early-warning signals that may indicate for a wide class of systems if a critical threshold is approaching"

Surprisingly, the notion of *Early Warning Signal(EWS)* or *Early Warning Sign* does not seem to appear in the IPCC glossary as of yet. Recently, Brovkin et al.[Bro+21, Box 1] write about EWS in their terminology introduction:

Early Warning Signals (EWS)

"[...] are quantitative indicators of the proximity of a system to a tipping point[Dak+08]. EWS apply mathematical principles of dynamical systems to Earth system components. EWS could be measured in one-dimensional space (such as time series of dust deposition in a marine core) using uni-variate precursors (for example, increasing temporal autocorrelation) or in multi-dimensional space (such as spatial patterns of vegetation cover) applying spatially explicit precursors" *(citation adapted)*

A lot of information on EWS is provided by [Dak+12]) on the EWS Tool Box Website [Dak+]. Dakos et al.[Dak+12] distinguish between *metric-based* and *model-based* methods for detecting EWS. Both kinds "reflect changes in the properties of the observed time-series" considered as generated by a stochastic dynamical system

> "*Metric-based* indicators quantify changes in the statistical properties of the time series [...] without attempting to fit the data with a specific model structure. *Model-based* methods quantify changes in the time series by attempting to fit the data to a model that is based on the general structure of" [*the generating model*]. "The ultimate goal of both types of indicators is to capture changes in the 'memory' (i.e. correlation structure) and variability of a time series and to determine if they follow patterns as predicted by models of critical transitions, while the system is approaching a transition into an alternative dynamic regime"

Caveats. Based on the example of the D-O events, Ditlevsen et al.[DJ10] point out two important provisos concerning the hope for detecting approaching abrupt changes by EWS. The first proviso concerns increased variance and increased autocorrelation as two generic characteristics indicating the approach of a bifurcation point:

> "These two signals are connected, and the detection of only one and not the other cannot be taken as a sign of an approaching tipping point. This is contrary to what was recently claimed [Dak+08; Sch+09]." *(citations adapted)*

The second proviso concerns the D-O events as "most clearly observed transition [...] besides the glacial-interglacial transitions themselves". Ditlevsen et al. investigate whether increased variance and increased autocorrelation can be detected in the NGRIP records [Mem04], arriving at the conclusion that D-O events are "most probably not generated by bifurcations: They are noise-induced transitions without early warning signals."
Subsequent research by Rypdal [Ryp16] and Boers [Boe18] however reports the presence of statistical EWS for D-O events. Yet it is important to note that not all abrupt changes – e.g. because they are noise- instead of bifurcation-induced – may be preceded by EWS. And also the possibility of false positives has to be taken into consideration. As Boers writes [Boe18]:

> "The fact that a forthcoming bifurcation implies the presence of EWS does, however, not exclude the possibility that statistical fluctuations indistinguishable from EWS arise either by chance or by other mechanisms unrelated to a bifurcation or abrupt transition."

And the authors of [Kéf+13] point out concerning EWS based on "critical slowing down":

> " slowing down generally happens in situations where a system is becoming increasingly sensitive to external perturbations, independent of whether the impeding change is catastrophic or not. These results highlight that indicators specific to catastrophic shifts are still lacking."

Livina, Ditlevsen and Lenton also address the criticism of [DJ10] by "blind testing" methods of detecting multiple system states and bifurcations in time series data [LDL12]: Time series data was provided by one author who knew the underlying model from which the data was generated, while the other two authors used different methods to analyse the data, “described the expected properties and behaviour of the underlying systems and attempted to model the corresponding data”. They conclude:

> “In most cases, the methods successfully detected the number of states in a system, and the occurrence of transitions between states. The derived models were able to reproduce the test data accurately. However, noise-induced abrupt transitions between existing states cannot be forecast due to the lack of any change in the underlying potential.”

Methods. We have seen that the notion *Early Warning Signal* refers to various statistical indicators for the approach of abrupt transitions in time series (often thought of as induced by bifurcations). A concrete step-by-step illustration of analysing time series data to detect approaching transitions is given in [Dak+12]. The data may come from proxy data or observations, or may have been generated by model simulation. Given a time-series as input, the procedure very roughly consists of

- pre-processing the data to obtain a *regular* time series, (i.e. there are no missing values and the data is equally spaced in time) without distorting its statistical properties
- filtering to eliminate non-stationarities in the mean of the time series (as they “can cause spurious indications of impending transitions”) and detrending to remove e.g. seasonal trends (since they would “impose a strong correlation structure on the time series”)
- apply one or (better) multiple methods to test the data for the presence of EWS.
  - metric-based: compute how certain statistical metrics for the time series, e.g. *variance*, *autocorrelation*, *skewness* [36] change along the time series;
  - model-based: fit the data to a specific kind of model, e.g. a *drift-diffusion-jump*, a *time-varying AR(p)* or a *threshold AR(p)* model. Properties of the fitted models then can be used as indicators.
- perform a sensitivity analysis by varying underlying choices of parameters in the previous steps, e.g. when using a *rolling window* method by altering the window size or the degree of smoothing for used in filtering.
- perform a significance test for the obtained results (whether EWS are considered to occur in the time series or not), especially to avoid false positives (type I errors); how this can be done depends on which method has been employed to test for EWS.

7.1 Selected references

- Wissel [Wis84] already studied indicators for approaching thresholds for ecosystems.
- Marotzke addresses the question predictability for a possible collapse of the THC [Mar00].
- Early papers suggesting data analysis methods to detect approaching thresholds: [KHP03; HK04; LL07]
- First papers explicitly using the term “EWS”: [Dak+08; Sch+09]
- Noise-induced abrupt transitions might not be preceded by EWS: [DJ10]; blind testing detection methods: [LDL12]
- There might also be false positives: EWS before “non-catastrophic” transitions [Kéf+13].

[36] i.e. how asymmetric the distribution of values of the time series is

- Kuehn suggests the mathematical theory of deterministic and stochastic *slow-fast* systems as providing a natural framework for studying critical transitions and EWS [Kue11; Kue13; Kue15] .

▸ **TiPES** Line of work leading to Objective 5.3

- Comparison and robustness testing of methods: [Len+12]
- Recent paper on using spectral analysis to detect more specific EWS that contain more information about the type of underlying bifurcation: [BBA20]
- EWS in paleo data: [LKL10; Ryp16; Boe18]

▸ **TiPES** Line of work leading to Objectives 1.1,1.2 and 1.4

8 Toward an ontology

Recapitulating the different objects, notions and computational patterns we have encountered in this report, we might consider a number of abstract concepts to organise them in the style of an ontology. In the following we sketch some ideas without being conclusive.
Consider the following two basic classes of objects (parameterised over types $\mathcal{I}, \mathcal{I}_1, \mathcal{I}_2, D, I, \mathit{Val}$):

- Sequential data that might be thought of as temporally ordered but potentially distributed in space (like *observation data, proxy records, trajectories of dynamical systems*)

  $$\mathit{SeqData}_{\mathcal{I},D} = \mathcal{I} \to D$$

  for a strictly ordered index type $\mathcal{I}$ and type of data D. An example of this could be a regular time series $\mathcal{T} \to X$ or the result of a numerical simulation $[x_0, \ldots, x_n]$ represented as function $\mathbb{N}_{<n} \to X$.

- Producers of sequential data (like *physical measurements* or *models of the climate system*) represented as functions that given some input I return a sequence of data:

  $$\mathit{SeqDataProducer}_{I,\mathcal{I},D} = I \to \mathit{SeqData}_{\mathcal{I},D}.$$

  In the case of measurements the "input" could be collected by sensors and associated to a timestamp to produce sequential data for further usage; for a dynamical system, the input could be an initial time and state, used to produce a trajectory as output.

Note that each *SeqData* object can be considered as a constant *SeqDataProducer*.
Given these two base classes of objects, we might consider the following classes of operations:

- Transformations of sequential data

  $$\mathit{SeqDataTr}_{\mathcal{I}_1,\mathcal{I}_2,D,\mathcal{T}} = \mathit{List}(\mathit{SeqData}_{\mathcal{I}_1,D}) \to \mathit{SeqData}_{\mathcal{I}_2,D}$$

  e.g. various operations on time series like *dating*, *removing biases* or *stacking*.

- Statistical analysis of sequential data

  $$\mathit{SeqDataStatAnalysis}_{\mathcal{I},D} = \mathit{SeqData}_{\mathcal{I},D} \to \mathit{Val}$$

  e.g. computing measures like variance, auto-correlation or simply the expected value.

- Comparison experiments in which, given one or more *SeqDataProducer* m_i and a list of inputs, the output produced by the m_i when applied to the different inputs is compared

  $$\mathit{ComparisonExperiment}_{I,\mathcal{I},D,\mathit{Val}} = \mathit{List}(\mathit{SeqDataProducer}_{I,\mathcal{I},D}) \times \mathit{List}(I) \to \mathit{Val}$$

  as e.g. seen in the computations of CS and commitment based on *comparisonExperiment* (Eq. (32)) for the special case in which there is only one *SeqDataProducer*, namely the climate model.

Other, more specific classes of experiments may be based on comparison experiments:

- Calibration experiments

  $$\begin{aligned}&\mathit{CalibrationExperiment}_{I,\mathcal{I},D,\mathit{Val}} =\\ &\qquad (P \to \mathit{SeqDataProducer}_{I,\mathcal{I},D}) \times \mathit{List}(P) \times I \times \mathit{List}(D) \to \mathit{Val}\end{aligned}$$

  in which, given a family of sequential data producers $\{m_p\}_{p:P}$, a list of possible parameter values $ps : \mathit{List}(P)$, an input and *SeqData* $d : D$ for validation, for example a ranking of the parameter values[37] is computed by performing comparison experiments between each instance of m_p (with p from ps) and d (considering d as constant *SeqDataProducer*). The CS-parameter experiment described in Section 4.1 is an example.

- Perturbation experiments

  $$\mathit{PerturbationExperiment}_{I,\mathcal{I},D,\mathit{Val}} = \mathit{SeqDataProducer}_{I,\mathcal{I},D} \times I \times (I \to I) \to \mathit{Val}$$

  in which, given a $\mathit{SeqDataProducer}_{I,\mathcal{I},D}$ m, an input $i : I$ and a function $\mathit{perturb} : I \to I$, a comparison experiment is performed with $[m, m]$ and $[i, \mathit{perturb}(i)]$ as inputs. The ECS experiment as instance of *perturbExperiment* in Section 4.1 is an example.

We have classes of properties

- for sequential data

  $$\mathit{SeqDataProp}_{\mathcal{I},D} = \mathit{SeqData}_{\mathcal{I},D} \to \mathit{Val}$$

- for sequential data producers

  $$\mathit{SeqDataProducerProp}_{I,\mathcal{I},D} = \mathit{SeqDataProducer}_{I,\mathcal{I},D} \to \mathit{Val}$$

where for logical properties $\mathit{Val} = \mathit{Prop}$ or $\mathit{Val} = \mathbb{B}$; for quantitative properties Val is a numerical type like $\mathbb{R}$, on which standard arithmetic operations are defined and which carries additional structure like an order $\sqsubseteq : \mathit{Val} \times \mathit{Val} \to \mathit{Prop}$ and a metric $\mathit{compare} : \mathit{Val} \times \mathit{Val} \to \mathbb{R}$.
To represent different kinds of uncertainties, we may consider monadic generalisations of the above, for example

$$\mathit{MSeqData}_{\mathcal{I},D,M} = \mathcal{I} \to M(D)$$

represented as functions that given some input I return a monadic value of data (e.g. a probability distribution). A generalised form of data producer could in turn produce monadic values of monadic sequential data:

$$\mathit{MSeqDataProducer}_{I,\mathcal{I},D,M} = I \to M(\mathit{MSeqData}_{\mathcal{I},D,M}).$$

where M is the functor of a monad.
In the examples mentioned above, we have already seen how the climate sensitivity and commitment experiments fit in this abstract framework. What about the other notions we have discussed?
We suggest that

- the different variants of climate sensitivity and climate change commitment discussed in this report are quantitative properties of sequential data producers;
- *having abrupt changes* or *having EWS* are logical properties of sequential data;
- *being a tipping element* and *having a tipping point* are logical properties of sequential data producers (which overlap in the Lenton et al. definition);
- *a tipping element* is a sequential data producer that has the above two properties and which moreover can produce data *having abrupt changes*. It is not necessary that it can produce sequential data *having EWS*.

[37] There are many options what exactly to compute as output: One may just compute one “best” parameter configuration according to some metric, or the best within configurations within a certain range etc.

9 Conclusion

Having reviewed a number of notions relevant for tipping point research, what have we learned? The object that one is ultimately interested in studying is the "real" climate system. As we have seen, climate sensitivity, climate change commitment and abrupt changes concern in a broad sense the response of the climate system or its subsystems to interventions:

- climate sensitivity is an idealised metric for the temperature increase resulting from a doubling of the CO_2 concentration of the atmosphere in an equilibrium state
- climate change commitment is a metric for delayed responses like temperature increase or sea level rise after forcing is stopped or at least not increased anymore in a non-equilibrium state
- abrupt change refers to a qualitative change of the system's behaviour in response to comparatively small changes in forcing.

Since the climate system is not amenable to systematic and repeatable empirical experiments, these properties can only be studied by analysing observations and studying representations of the system or its subsystems. For this, time series and dynamical systems play a crucial role, and there are generic techniques that are used to study their properties. What we called *comparison experiment* in Sections 4 and 5 is such a generic technique, and similarly the fitting of parameterised models to time series of observations. The methods that are used to study the notions we discussed in this report arise as instances of such generic techniques, and the notions might be studied in more than one way, as we have e.g. seen for climate sensitivity. In the preparation of this report and from discussions in the context of the TiPES project, our impression was that it is difficult to attach unambiguous "meanings" to these notions. This has been noted before. Russill [Rus15] reviews origins, precursors, and debates around the usage of the "tipping point" in the discourse about climate change. He points out that besides the framework of dynamical systems theory, the usage of tipping point in the climate literature is also related to the scientific community inspired by the Gaia theory [Lov72]. Still following [Rus15], the tipping point is a metaphor that has the property of being "generative": its cultural resonance is intended to "shift popular heuristics for environmental change". However, the multiple usages of the "tipping point", stemming from different communities, come at the cost of semantic confusion (see below).

For the notions of abrupt change, TP and EWS, in particular, the usages seem to be evolving. In reviewing the literature, our impression was that descriptions in the first papers on abrupt climate change, tipping and EWS seem inspired by bifurcations in dynamical systems without necessarily saying this explicitly [RN09]. In the line of work by Lenton et al. based on [Len+08], TPs are however understood in a more liberal way, as seen in Section 6.4. This is e.g. discussed in [Len12, pp.11-12] and, again, disagreements arising from different perceptions of what is meant by TP are called *semantic confusion* in [Dua+12]. Similarly, following [Dak+08], the term EWS sometimes seemed to be used as a synonym for *critical slowing down* exhibited by dynamical systems when approaching a bifurcation [DJ10]. But other indicators, also for different kinds of tipping, have been proposed and also called EWS [Dak+12; RS16; Ma+22].

For this report, we have collected informal definitions and classification information that might be used as semantic annotation for different instances of climate sensitivity and commitment experiments, tipping elements and early warning signs. We have furthermore suggested generic schemes of which CS and commitment experiments are instances and discussed Lenton et al.'s formal definition of TEs. Finally, we have suggested an abstract perspective that may help to organise tipping notions in the style of an ontology.

Acknowledgements

This report is part of TiPES Deliverable 6.2. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 820970 (TiPES –Tipping Points in the Earth System, 2019–2023).

We thank our TiPES colleague Robbin Bastiaansen for helpful discussions about climate sensitivity. Moreover, the TiPES Gathertown meetings organised by Robbin together with Keno Riechers provided a nice opportunity for discussions around various topics relevant for this report. We furthermore thank our external collaborator Patrik Jansson for his valuable advice during the preparation of this report. Many thanks also go to our colleague Tim Richter for long discussions about the material and careful proof-reading of a preliminary version of the manuscript.

The work presented in this paper heavily relies on free software, among others on Idris, git, Emacs, LaTeX and on the FreeBSD and Debian GNU/Linux operating systems. It is our pleasure to thank all developers of these excellent products.

Conflicts of Interest

None.

General References

[GL20] Michael Ghil and Valerio Lucarini. "The physics of climate variability and climate change". In: *Reviews of Modern Physics* 92.3 (2020), p. 035002.

[Goo+10] Hugues Goosse et al. *Introduction to climate dynamics and climate modeling.* Centre de recherche sur la Terre et le climat Georges Lemaître-UCLouvain, 2010.

[Goo15] Hugues Goosse. *Climate system dynamics and modeling.* Cambridge University Press, 2015.

[Sto11] Thomas Stocker. *Introduction to climate modelling.* Springer Science & Business Media, 2011.

IPCC References

[IPC18] IPCC. "Annex I: Glossary in Global Warming of 1.5C. An IPCC Special Report on the impacts of global warming of 1.5C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the thre". In: (2018).

[Kri+13] Gerhard Krinner et al. "Long-term climate change: Projections, commitments and irreversibility". In: *Climate Change 2013 the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* 9781107057999 (2013), pp. 1029–1136. DOI: 10.1017/CBO9781107415324.024.

[Mas+21] V. Masson-Delmotte et al., eds. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (in press).* Cambridge University Press, 2021.

[Mat+21] J.B.R. Matthews et al. "IPCC 2021: Annex VII: Glossary (in press)". In: *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (in press)* (2021).

[Mee+07] Gerald A Meehl et al. *Global climate projections. Climate change 2007: the physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.* 2007.

References: Models

[Bud69] Mikhail I Budyko. “The effect of solar radiation variations on the climate of the Earth”. In: *tellus* 21.5 (1969), pp. 611–619.

[Fla11] Gregory M Flato. “Earth system models: an overview”. In: *Wiley Interdisciplinary Reviews: Climate Change* 2.6 (2011), pp. 783–800.

[Fra79] Klaus Fraedrich. “Catastrophes and resilience of a zero-dimensional climate system with ice-albedo and greenhouse feedback”. In: *Quarterly Journal of the Royal Meteorological Society* 105.443 (1979), pp. 147–167.

[Lor63] Edward N Lorenz. “Deterministic nonperiodic flow”. In: *Journal of atmospheric sciences* 20.2 (1963), pp. 130–141.

[Mar00] Jochem Marotzke. “Abrupt climate change and thermohaline circulation: Mechanisms and predictability”. In: *Proceedings of the National Academy of Sciences* 97.4 (2000), pp. 1347–1350.

[Roo82] Claes Rooth. “Hydrology and ocean circulation”. In: *Progress in Oceanography* 11.2 (1982), pp. 131–149.

[Sel69] William D Sellers. “A global climatic model based on the energy balance of the earth-atmosphere system”. In: *Journal of Applied Meteorology and Climatology* 8.3 (1969), pp. 392–400.

[Sto61] Henry Stommel. “Thermohaline convection with two stable regimes of flow”. In: *Tellus* 13.2 (1961), pp. 224–230.

[Sut81] Alfonso Sutera. “On stochastic perturbation and long-term climate behaviour”. In: *Quarterly Journal of the Royal Meteorological Society* 107.451 (1981), pp. 137–151.

References: Scenarios

[She+18] Theodore G Shepherd et al. “Storylines: an alternative approach to representing uncertainty in physical aspects of climate change”. In: *Climatic change* 151.3 (2018), pp. 555–571.

References: Climate Sensitivity

[AGW15] Timothy Andrews, Jonathan M Gregory, and Mark J Webb. “The dependence of radiative forcing and feedback on evolving patterns of surface temperature change in climate models”. In: *Journal of Climate* 28.4 (2015), pp. 1630–1648.

[Arr96] Svante Arrhenius. “On the influence of carbonic acid in the air upon the temperature of the ground”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 41.251 (1896), pp. 237–276.

[Bud69] Mikhail I Budyko. “The effect of solar radiation variations on the climate of the Earth”. In: *tellus* 21.5 (1969), pp. 611–619.

[Cha+79] Jule G Charney et al. *Carbon dioxide and climate: a scientific assessment.* 1979.

[Gre+04] J. M. Gregory et al. “A new method for diagnosing radiative forcing and climate sensitivity”. In: *Geophysical Research Letters* 31.3 (2004). DOI: 10.1029/2003GL018747. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2003GL018747. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2003GL018747.

[Han+05] James Hansen et al. “Efficacy of climate forcings”. In: *Journal of Geophysical Research: Atmospheres* 110.D18 (2005).

[Han+84] J Hansen et al. “Climate sensitivity: Analysis of feedback mechanisms.” In: *feedback* 1 (1984), pp. 1–3.

[Hey+16] Anna S von der Heydt et al. "Lessons on climate sensitivity from past climate changes". In: *Current Climate Change Reports* 2.4 (2016), pp. 148–158.

[KR15] Reto Knutti and Maria AA Rugenstein. "Feedbacks, climate sensitivity and the limits of linear models". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 373.2054 (2015), p. 20150146.

[KRH17] Reto Knutti, Maria AA Rugenstein, and Gabriele C Hegerl. "Beyond equilibrium climate sensitivity". In: *Nature Geoscience* 10.10 (2017), pp. 727–736.

[Kue13] Christian Kuehn. "A mathematical framework for critical transitions: normal forms, variance and applications". In: *Journal of Nonlinear Science* 23.3 (2013), pp. 457–510.

[MW75] Syukuro Manabe and Richard T Wetherald. "The effects of doubling the CO2 concentration on the climate of a general circulation model". In: *Journal of Atmospheric Sciences* 32.1 (1975), pp. 3–15.

[Myh+98] Gunnar Myhre et al. "New estimates of radiative forcing due to well mixed greenhouse gases". In: *Geophysical research letters* 25.14 (1998), pp. 2715–2718.

[Roh+12] Eelco J Rohling et al. "Making sense of palaeoclimate sensitivity". In: *Nature* 491 (2012), pp. 683–691.

[Sch83] Michael E Schlesinger. "A review of climate models and their simulation of CO2-induced warming". In: *International Journal of Environmental Studies* 20.2 (1983), pp. 103–114.

[Sel69] William D Sellers. "A global climatic model based on the energy balance of the earth-atmosphere system". In: *Journal of Applied Meteorology and Climatology* 8.3 (1969), pp. 392–400.

[She+20] SC Sherwood et al. "An assessment of Earth's climate sensitivity using multiple lines of evidence". In: *Reviews of Geophysics* 58.4 (2020), e2019RG000678.

[Ste+16] Bjorn Stevens et al. "Prospects for narrowing bounds on Earth's equilibrium climate sensitivity". In: *Earth's Future* 4.11 (2016), pp. 512–522.

References: Commitment

[CC10] Long Cao and Ken Caldeira. "Atmospheric carbon dioxide removal: long-term consequences and commitment". In: *Environmental Research Letters* 5.2 (2010), p. 024011.

[Eby+09] M Eby et al. "Lifetime of anthropogenic climate change: Millennial time scales of potential CO2 and surface temperature perturbations". In: *Journal of climate* 22.10 (2009), pp. 2501–2511.

[FP09] Christophe Feltus and Michaël Petit. "Building a responsibility model using modal logic-towards Accountability, Aapability and Commitment concepts". In: *2009 IEEE/ACS International Conference on Computer Systems and Applications*. IEEE. 2009, pp. 386–391.

[Han+02] J Hansen et al. "Climate forcings in Goddard Institute for space studies SI2000 simulations". In: *Journal of Geophysical Research: Atmospheres* 107.D18 (2002), ACL–2.

[Han+05] James Hansen et al. "Efficacy of climate forcings". In: *Journal of Geophysical Research: Atmospheres* 110.D18 (2005).

[Han+85] James Hansen et al. "Climate response times: Dependence on climate sensitivity and ocean mixing". In: *Science* 229.4716 (1985), pp. 857–859.

[Hel+10] Isaac M Held et al. "Probing the fast and slow components of global warming by returning abruptly to preindustrial forcing". In: *Journal of Climate* 23.9 (2010), pp. 2418–2427.

[Kri+13] Gerhard Krinner et al. "Long-term climate change: Projections, commitments and irreversibility". In: *Climate Change 2013 the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* 9781107057999 (2013), pp. 1029–1136. DOI: 10.1017/CBO9781107415324.024.

[Mee+07] Gerald A Meehl et al. *Global climate projections. Climate change 2007: the physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.* 2007.

[Pat+18] Frank Pattyn et al. "The Greenland and Antarctic ice sheets under 1.5 °C global warming". In: *Nature Climate Change* 8.12 (2018), pp. 1053–1061. ISSN: 17586798. DOI: 10.1038/s41558-018-0305-8. URL: http://dx.doi.org/10.1038/s41558-018-0305-8.

[Pie03] Roger A Pielke Sr. "Heat storage within the Earth system". In: *Bulletin of the American Meteorological Society* 84.3 (2003), pp. 331–336.

[Pla+08] Gian Kasper Plattner et al. "Long-term climate commitments projected with climate-carbon cycle models". In: *Journal of Climate* 21.12 (2008), pp. 2721–2751. ISSN: 08948755. DOI: 10.1175/2007JCLI1905.1.

[Ram88] Veerabhachan Ramanathan. "The greenhouse theory of climate change: a test by an inadvertent global experiment". In: *Science* 240.4850 (1988), pp. 293–299.

[SM99] Ronald J Stouffer and Syukuro Manabe. "Response of a coupled ocean–atmosphere model to increasing atmospheric carbon dioxide: Sensitivity to the rate of increase". In: *Journal of Climate* 12.8 (1999), pp. 2224–2237.

[Sol+09] Susan Solomon et al. "Irreversible climate change due to carbon dioxide emissions". In: *Proceedings of the national academy of sciences* 106.6 (2009), pp. 1704–1709.

[Sto04] Ronald J Stouffer. "Time scales of climate response". In: *Journal of Climate* 17.1 (2004), pp. 209–217.

[Wig05] T. M. L. Wigley. "The Climate Change Commitment". In: *Science* 307.5716 (2005), pp. 1766–1769. ISSN: 0036-8075. DOI: 10.1126/science.1103934. eprint: https://science.sciencemag.org/content/307/5716/1766.full.pdf. URL: https://science.sciencemag.org/content/307/5716/1766.

[WR01] Tom ML Wigley and Sarah CB Raper. "Interpretation of high projections for global-mean warming". In: *Science* 293.5529 (2001), pp. 451–454.

[WR93] TML Wigley and SCB Raper. "Future changes in global mean temperature and sea level". In: *Climate and sea level change: observations, projections and implications* 111 (1993), p. 133.

[WSD01] Richard T Wetherald, Ronald J Stouffer, and Keith W Dixon. "Committed warming and its implications for climate change". In: *Geophysical Research Letters* 28.8 (2001), pp. 1535–1538.

References: Stability

[Roo82] Claes Rooth. "Hydrology and ocean circulation". In: *Progress in Oceanography* 11.2 (1982), pp. 131–149.

[Sto61] Henry Stommel. "Thermohaline convection with two stable regimes of flow". In: *Tellus* 13.2 (1961), pp. 224–230.

References: Abrupt change

[All+03] Richard B Alley et al. "Abrupt climate change". In: *science* 299.5615 (2003), pp. 2005–2010.

[All00] Richard B Alley. “Ice-core evidence of abrupt climate changes”. In: *Proceedings of the National Academy of Sciences* 97.4 (2000), pp. 1331–1334.

[Ank+93] Martin Anklin et al. “Climate instability during the last interglacial period recorded in the GRIP ice core”. In: *Nature* 364 (1993), pp. 203–207.

[Bar+11] Stephen Barker et al. “800,000 years of abrupt climate variability”. In: *science* 334.6054 (2011), pp. 347–351.

[BC+02] Ocean Studies Board, National Research Council, et al. *Abrupt Climate Change: Inevitable Surprises.* National Academies Press, 2002.

[BD89] Wallace S Broecker and George H Denton. “The role of ocean-atmosphere reorganizations in glacial cycles”. In: *Geochimica et Cosmochimica Acta* 53.10 (1989), pp. 2465–2501.

[BHD10] Edouard Bard, Bruno Hamelin, and Doriane Delanghe-Sabatier. “Deglacial meltwater pulse 1B and Younger Dryas sea levels revisited with boreholes at Tahiti”. In: *Science* 327.5970 (2010), pp. 1235–1237.

[Bon+93] Gerard Bond et al. “Correlations between climate records from North Atlantic sediments and Greenland ice”. In: *Nature* 365.6442 (1993), pp. 143–147.

[Bon+97] Gerard Bond et al. “A pervasive millennial-scale cycle in North Atlantic Holocene and glacial climates”. In: *science* 278.5341 (1997), pp. 1257–1266.

[Bro+21] Victor Brovkin et al. “Past abrupt changes, tipping points and cascading impacts in the Earth system”. In: *Nature Geoscience* (2021), pp. 1–9.

[Bro+92] Wallace Broecker et al. “Origin of the northern Atlantic’s Heinrich events”. In: *Climate Dynamics* 6.3 (1992), pp. 265–273.

[Bro02] W S Broecker. *The Glacial World According to Wally.* 3rd. Eldigio Press, 2002.

[Dan+84] Willi Dansgaard et al. “North Atlantic climatic oscillations revealed by deep Greenland ice cores”. In: *Climate processes and climate sensitivity* 29 (1984), pp. 288–298.

[Dan+93] Willi Dansgaard et al. “Evidence for general instability of past climate from a 250-kyr ice-core record”. In: *nature* 364.6434 (1993), pp. 218–220.

[DAS07] Peter D Ditlevsen, Katrine Krogh Andersen, and Anders Svensson. “The DO-climate events are probably noise induced: statistical investigation of the claimed 1470 years cycle”. In: *Climate of the Past* 3.1 (2007), pp. 129–134.

[Dem+00] Peter Demenocal et al. “Abrupt onset and termination of the African Humid Period:: rapid climate responses to gradual insolation forcing”. In: *Quaternary science reviews* 19.1-5 (2000), pp. 347–361.

[Dri+15] Sybren Drijfhout et al. “Catalogue of abrupt shifts in Intergovernmental Panel on Climate Change climate models”. In: *Proceedings of the National Academy of Sciences* 112.43 (2015), E5777–E5786.

[Hei88] Hartmut Heinrich. “Origin and consequences of cyclic ice rafting in the northeast Atlantic Ocean during the past 130,000 years”. In: *Quaternary research* 29.2 (1988), pp. 142–152.

[Lor63] Edward N Lorenz. “Deterministic nonperiodic flow”. In: *Journal of atmospheric sciences* 20.2 (1963), pp. 130–141.

[Mar+14] Shaun A Marcott et al. “Centennial-scale changes in the global carbon cycle during the last deglaciation”. In: *Nature* 514.7524 (2014), pp. 616–619.

[Mar00] Jochem Marotzke. “Abrupt climate change and thermohaline circulation: Mechanisms and predictability”. In: *Proceedings of the National Academy of Sciences* 97.4 (2000), pp. 1347–1350.

[Oes+84] Hans Oeschger et al. “Late glacial climate history from ice cores”. In: *Climate processes and climate sensitivity* 29 (1984), pp. 299–306.

[Sev+98] Jeffrey P Severinghaus et al. “Timing of abrupt climate change at the end of the Younger Dryas interval from thermally fractionated gases in polar ice”. In: *Nature* 391.6663 (1998), pp. 141–146.

[Sto00] Thomas F Stocker. “Past and future reorganizations in the climate system”. In: *Quaternary Science Reviews* 19.1-5 (2000), pp. 301–319.

[WFR09] EW Wolff, Hubertus Fischer, and R Röthlisberger. “Glacial terminations as southern warmings without northern control”. In: *Nature Geoscience* 2.3 (2009), pp. 206–209.

References: Tipping

[ABS16] Ori Adam, Tobias Bischoff, and Tapio Schneider. “Seasonal and interannual variations of the energy flux equator and ITCZ. Part I: Zonally averaged ITCZ position”. In: *Journal of Climate* 29.9 (2016), pp. 3219–3230.

[Alk+19] Hassan Alkhayuon et al. “Basin bifurcations, oscillatory instability and rate-induced thresholds for Atlantic meridional overturning circulation in a global oceanic box model”. In: *Proceedings of the Royal Society A* 475.2225 (2019), p. 20190051.

[Arm+] David Armstrong McKay et al. “Updated assessment of climate tipping elements suggests¿ 1.5 oC global warming could trigger multiple tipping points”. In: ().

[Ash+12] Peter Ashwin et al. “Tipping points in open systems: bifurcation, noise-induced and rate-dependent examples in the climate system”. In: *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* 370.1962 (2012), pp. 1166–1184. ISSN: 1364503X. URL: http://www.jstor.org/stable/41348437.

[ATW21] Hassan Alkhayuon, Rebecca C Tyson, and Sebastian Wieczorek. “Phase-sensitive tipping: How cyclic ecosystems respond to contemporary climate”. In: *arXiv preprint arXiv:2101.12107* (2021).

[Dua+12] Carlos M Duarte et al. “Abrupt climate change in the Arctic”. In: *Nature Climate Change* 2.2 (2012), pp. 60–62.

[Fra79] Klaus Fraedrich. “Catastrophes and resilience of a zero-dimensional climate system with ice-albedo and greenhouse feedback”. In: *Quarterly Journal of the Royal Meteorological Society* 105.443 (1979), pp. 147–167.

[Gin+18] Kees van Ginkel et al. “D3. 1 Operationalizing socio-economic and climate tipping points”. In: *Deliverable of the H2020 COACCH project* (2018).

[Klo+21] Ann Kristin Klose et al. “What do we mean,‘tipping cascade’?” In: *Environmental Research Letters* 16.12 (2021), p. 125011.

[Krö+20] Jonathan Krönke et al. “Dynamics of tipping cascades on complex networks”. In: *Physical Review E* 101.4 (2020), p. 042311.

[Kue13] Christian Kuehn. “A mathematical framework for critical transitions: normal forms, variance and applications”. In: *Journal of Nonlinear Science* 23.3 (2013), pp. 457–510.

[Kue15] Christian Kuehn. *Multiple time scale dynamics*. Vol. 191. Springer, 2015.

[Len+08] Timothy M Lenton et al. “Tipping elements in the Earth’s climate system”. In: *Proceedings of the national Academy of Sciences* 105.6 (2008), pp. 1786–1793.

[Len+19] Timothy Lenton et al. “Climate tipping points - too risky to bet against”. In: *Nature* 575 (2019), pp. 592–595.

[Len12] Timothy M Lenton. “Arctic climate tipping points”. In: *Ambio* 41.1 (2012), pp. 10–22.

[Len13] Timothy M Lenton. “Environmental tipping points”. In: *Annual Review of Environment and Resources* 38 (2013), pp. 1–29.

[Len20] Timothy M Lenton. “Tipping positive change”. In: *Philosophical Transactions of the Royal Society B* 375.1794 (2020), p. 20190123.

[Loh+21] Johannes Lohmann et al. "Abrupt climate change as a rate-dependent cascading tipping point". In: *Earth System Dynamics* 12.3 (2021), pp. 819–835.

[LS07] Timothy M Lenton and Hans Joachim Schellnhuber. "Tipping the scales". In: *Nature Climate Change* 1.712 (2007), pp. 97–98.

[McK+] David I Armstrong McKay et al. "Updated assessment suggests¿ 1.5 C global warming could trigger multiple climate tipping points". In: ().

[RN09] Chris Russill and Zoe Nyssa. "The tipping point trend in climate change communication". In: *Global environmental change* 19.3 (2009), pp. 336–344.

[Rus11] Chris Russill. "Temporal metaphor in abrupt climate change communication: an initial effort at clarification". In: *The economic, social and political elements of climate change.* Springer, 2011, pp. 113–132.

[Rus15] Chris Russill. "Climate change tipping points: origins, precursors, and debates". In: *Wiley Interdisciplinary Reviews: Climate Change* 6.4 (2015), pp. 427–434.

[SL21] Simon Sharpe and Timothy M Lenton. "Upward-scaling tipping cascades to meet climate goals: Plausible grounds for hope". In: *Climate Policy* 21.4 (2021), pp. 421–433.

[Sut81] Alfonso Sutera. "On stochastic perturbation and long-term climate behaviour". In: *Quarterly Journal of the Royal Meteorological Society* 107.451 (1981), pp. 137–151.

[Trö+17] Jenny Tröltzsch et al. "D1.2 Knowledge synthesis and gap analysis on climate impact analysis, economic costs and scenarios". In: (2017).

[VHS18] Sandra Van der Hel, Iina Hellsten, and Gerard Steen. "Tipping points and climate change: Metaphor between science and the media". In: *Environmental Communication* 12.5 (2018), pp. 605–620.

[Wie+11] S. Wieczorek et al. "Excitability in ramped systems: the compost-bomb instability". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 467.2129 (2011), pp. 1243–1269. DOI: 10.1098/rspa.2010.0485.

[Win+20] Ricarda Winkelmann et al. "Social tipping processes for sustainability: An analytical framework". In: *arXiv preprint arXiv:2010.04488* (2020).

[Wun+21] Nico Wunderling et al. "Interacting tipping elements increase risk of climate domino effects under global warming". In: *Earth System Dynamics* 12.2 (2021), pp. 601–619.

References: Early Warning Signals

[BBA20] Thomas M Bury, Chris T Bauch, and Madhur Anand. "Detecting and distinguishing tipping points using spectral early warning signals". In: *Journal of the Royal Society Interface* 17.170 (2020), p. 20200482.

[Boe18] Niklas Boers. "Early-warning signals for Dansgaard-Oeschger events in a high-resolution ice core record". In: *Nature communications* 9.1 (2018), pp. 1–8.

[Dak+] V. Dakos et al. *Early Warning Signals Toolbox*. https://www.early-warning-signals.org/, last accessed 22/04/13.

[Dak+08] Vasilis Dakos et al. "Slowing down as an early warning signal for abrupt climate change". In: *Proceedings of the National Academy of Sciences* 105.38 (2008), pp. 14308–14312.

[Dak+12] Vasilis Dakos et al. "Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data". In: *PloS one* 7.7 (2012), e41010.

[DJ10] Peter D Ditlevsen and Sigfus J Johnsen. "Tipping points: Early warning and wishful thinking". In: *Geophysical Research Letters* 37.19 (2010).

[Kéf+13] Sonia Kéfi et al. "Early warning signals also precede non-catastrophic transitions". In: *Oikos* 122.5 (2013), pp. 641–648.

[KHP03] Thomas Kleinen, Hermann Held, and Gerhard Petschel-Held. “The potential role of spectral properties in detecting thresholds in the Earth system: application to the thermohaline circulation”. In: *Ocean Dynamics* 53.2 (2003), pp. 53–63.

[KLN22] Christian Kuehn, Kerstin Lux, and Alexandra Neamţu. “Warning signs for non-Markovian bifurcations: colour blindness and scaling laws”. In: *Proceedings of the Royal Society A* 478.2259 (2022), p. 20210740.

[KS02] Reto Knutti and Thomas F Stocker. “Limited predictability of the future thermohaline circulation close to an instability threshold”. In: *Journal of Climate* 15.2 (2002), pp. 179–186.

[Kue15] Christian Kuehn. *Multiple time scale dynamics.* Vol. 191. Springer, 2015.

[LDL12] VN Livina, PD Ditlevsen, and TM Lenton. “An independent test of methods of detecting system states and bifurcations in time-series data”. In: *Physica A: Statistical Mechanics and its Applications* 391.3 (2012), pp. 485–496.

[Len+12] TM Lenton et al. “Early warning of climate tipping points from critical slowing down: comparing methods to improve robustness”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 370.1962 (2012), pp. 1185–1204.

[LL07] Valerie N Livina and Timothy M Lenton. “A modified method for detecting incipient bifurcations in a dynamical system”. In: *Geophysical research letters* 34.3 (2007).

[Mar00] Jochem Marotzke. “Abrupt climate change and thermohaline circulation: Mechanisms and predictability”. In: *Proceedings of the National Academy of Sciences* 97.4 (2000), pp. 1347–1350.

[Mem04] North GRIP Members. “High resolution Climate Record of the Northern Hemisphere reaching into the last Glacial Interglacial Period”. In: *Nature* 431 (2004), pp. 147–151.

[Ryp16] Martin Rypdal. “Early-warning signals for the onsets of Greenland interstadials and the Younger Dryas–Preboreal transition”. In: *Journal of Climate* 29.11 (2016), pp. 4047–4056.

[Sch+09] Marten Scheffer et al. “Early-warning signals for critical transitions”. In: *Nature* 461.7260 (2009), pp. 53–59.

[Tzi00] Eli Tziperman. “Proximity of the present-day thermohaline circulation to an instability threshold”. In: *Journal of Physical Oceanography* 30.1 (2000), pp. 90–104.

[Wis84] C Wissel. “A universal law of the characteristic return time near thresholds”. In: *Oecologia* 65.1 (1984), pp. 101–107.

All References

[23] *TiPES H2020 Project Website*. https://www.tipes.dk/. 2019–2023.

[ABS16] Ori Adam, Tobias Bischoff, and Tapio Schneider. “Seasonal and interannual variations of the energy flux equator and ITCZ. Part I: Zonally averaged ITCZ position”. In: *Journal of Climate* 29.9 (2016), pp. 3219–3230.

[AGW15] Timothy Andrews, Jonathan M Gregory, and Mark J Webb. “The dependence of radiative forcing and feedback on evolving patterns of surface temperature change in climate models”. In: *Journal of Climate* 28.4 (2015), pp. 1630–1648.

[Alk+19] Hassan Alkhayuon et al. “Basin bifurcations, oscillatory instability and rate-induced thresholds for Atlantic meridional overturning circulation in a global oceanic box model”. In: *Proceedings of the Royal Society A* 475.2225 (2019), p. 20190051.

[All+03] Richard B Alley et al. “Abrupt climate change”. In: *science* 299.5615 (2003), pp. 2005–2010.

[All00] Richard B Alley. “Ice-core evidence of abrupt climate changes”. In: *Proceedings of the National Academy of Sciences* 97.4 (2000), pp. 1331–1334.

[Ank+93] Martin Anklin et al. “Climate instability during the last interglacial period recorded in the GRIP ice core”. In: *Nature* 364 (1993), pp. 203–207.

[Arm+] David Armstrong McKay et al. “Updated assessment of climate tipping elements suggests¿ 1.5 oC global warming could trigger multiple tipping points”. In: ().

[Arn+13] Vladimir Igorevich Arnold et al. *Dynamical systems V: bifurcation theory and catastrophe theory*. Vol. 5. Springer Science & Business Media, 2013.

[Arr96] Svante Arrhenius. “On the influence of carbonic acid in the air upon the temperature of the ground”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 41.251 (1896), pp. 237–276.

[Ash+12] Peter Ashwin et al. “Tipping points in open systems: bifurcation, noise-induced and rate-dependent examples in the climate system”. In: *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* 370.1962 (2012), pp. 1166–1184. ISSN: 1364503X. URL: http://www.jstor.org/stable/41348437.

[ATW21] Hassan Alkhayuon, Rebecca C Tyson, and Sebastian Wieczorek. “Phase-sensitive tipping: How cyclic ecosystems respond to contemporary climate”. In: *arXiv preprint arXiv:2101.12107* (2021).

[Bar+11] Stephen Barker et al. “800,000 years of abrupt climate variability”. In: *science* 334.6054 (2011), pp. 347–351.

[Bar+84] Hendrik P Barendregt et al. *The lambda calculus*. Vol. 3. North-Holland Amsterdam, 1984.

[Bar+96] G.I. Barenblatt et al. *Scaling, Self-similarity, and Intermediate Asymptotics: Dimensional Analysis and Intermediate Asymptotics*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 1996. ISBN: 9780521435222. URL: https://books.google.de/books?id=r-Az53e-MTYC.

[BBA20] Thomas M Bury, Chris T Bauch, and Madhur Anand. “Detecting and distinguishing tipping points using spectral early warning signals”. In: *Journal of the Royal Society Interface* 17.170 (2020), p. 20200482.

[BC+02] Ocean Studies Board, National Research Council, et al. *Abrupt Climate Change: Inevitable Surprises*. National Academies Press, 2002.

[BD89] Wallace S Broecker and George H Denton. “The role of ocean-atmosphere reorganizations in glacial cycles”. In: *Geochimica et Cosmochimica Acta* 53.10 (1989), pp. 2465–2501.

[BHD10] Edouard Bard, Bruno Hamelin, and Doriane Delanghe-Sabatier. “Deglacial meltwater pulse 1B and Younger Dryas sea levels revisited with boreholes at Tahiti”. In: *Science* 327.5970 (2010), pp. 1235–1237.

[Boe18] Niklas Boers. “Early-warning signals for Dansgaard-Oeschger events in a high-resolution ice core record”. In: *Nature communications* 9.1 (2018), pp. 1–8.

[Bon+93] Gerard Bond et al. “Correlations between climate records from North Atlantic sediments and Greenland ice”. In: *Nature* 365.6442 (1993), pp. 143–147.

[Bon+97] Gerard Bond et al. “A pervasive millennial-scale cycle in North Atlantic Holocene and glacial climates”. In: *science* 278.5341 (1997), pp. 1257–1266.

[Bot+21] Nicola Botta et al. “Responsibility Under Uncertainty: Which Climate Decisions Matter Most?” In: (2021).

[Bot21] Nicola Botta. *IdrisLibs*. https://gitlab.pik-potsdam.de/botta/IdrisLibs. 2016–2021.

[Bri22] P. W. Bridgman. *Dimensional Analysis*. Yale University Press, 1922.

[Bro+21] Victor Brovkin et al. "Past abrupt changes, tipping points and cascading impacts in the Earth system". In: *Nature Geoscience* (2021), pp. 1–9.

[Bro+92] Wallace Broecker et al. "Origin of the northern Atlantic's Heinrich events". In: *Climate Dynamics* 6.3 (1992), pp. 265–273.

[Bro02] W S Broecker. *The Glacial World According to Wally.* 3rd. Eldigio Press, 2002.

[Buc14] Edgar Buckingham. "On Physically Similar Systems; Illustrations of the Use of Dimensional Equations". In: *Phys . Rev.* 4.4 (1914), pp. 345–376.

[Buc15] Edgar Buckingham. "Model experiments and the form of physical equations". In: *Transactions of The American Society of Mechanical Engineers* 37 (1915), pp. 263–296.

[Bud69] Mikhail I Budyko. "The effect of solar radiation variations on the climate of the Earth". In: *tellus* 21.5 (1969), pp. 611–619.

[CC10] Long Cao and Ken Caldeira. "Atmospheric carbon dioxide removal: long-term consequences and commitment". In: *Environmental Research Letters* 5.2 (2010), p. 024011.

[Cha+79] Jule G Charney et al. *Carbon dioxide and climate: a scientific assessment.* 1979.

[Com86] NASA Advisory Council. Earth System Sciences Committee. *Earth system science overview: a program for global change.* National Aeronautics and Space Administration, 1986.

[Com88] NASA Advisory Council. Earth System Sciences Committee. *Earth system science: A closer view.* National Academies, 1988.

[Cru12] Michel Crucifix. "Traditional and novel approaches to palaeoclimate modelling". In: *Quaternary Science Reviews* 57 (2012), pp. 1–16. ISSN: 0277-3791. DOI: https://doi.org/10.1016/j.quascirev.2012.09.010. URL: https://www.sciencedirect.com/science/article/pii/S0277379112003472.

[Dak+] V. Dakos et al. *Early Warning Signals Toolbox.* https://www.early-warning-signals.org/, last accessed 22/04/13.

[Dak+08] Vasilis Dakos et al. "Slowing down as an early warning signal for abrupt climate change". In: *Proceedings of the National Academy of Sciences* 105.38 (2008), pp. 14308–14312.

[Dak+12] Vasilis Dakos et al. "Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data". In: *PloS one* 7.7 (2012), e41010.

[Dan+82] Willi Dansgaard et al. "A new Greenland deep ice core". In: *Science* 218.4579 (1982), pp. 1273–1277.

[Dan+84] Willi Dansgaard et al. "North Atlantic climatic oscillations revealed by deep Greenland ice cores". In: *Climate processes and climate sensitivity* 29 (1984), pp. 288–298.

[Dan+93] Willi Dansgaard et al. "Evidence for general instability of past climate from a 250-kyr ice-core record". In: *nature* 364.6434 (1993), pp. 218–220.

[DAS07] Peter D Ditlevsen, Katrine Krogh Andersen, and Anders Svensson. "The DO-climate events are probably noise induced: statistical investigation of the claimed 1470 years cycle". In: *Climate of the Past* 3.1 (2007), pp. 129–134.

[Dem+00] Peter Demenocal et al. "Abrupt onset and termination of the African Humid Period:: rapid climate responses to gradual insolation forcing". In: *Quaternary science reviews* 19.1-5 (2000), pp. 347–361.

[DJ10] Peter D Ditlevsen and Sigfus J Johnsen. "Tipping points: Early warning and wishful thinking". In: *Geophysical Research Letters* 37.19 (2010).

[Dri+15] Sybren Drijfhout et al. "Catalogue of abrupt shifts in Intergovernmental Panel on Climate Change climate models". In: *Proceedings of the National Academy of Sciences* 112.43 (2015), E5777–E5786.

[Dua+12] Carlos M Duarte et al. "Abrupt climate change in the Arctic". In: *Nature Climate Change* 2.2 (2012), pp. 60–62.

[Eby+09] M Eby et al. “Lifetime of anthropogenic climate change: Millennial time scales of potential CO2 and surface temperature perturbations”. In: *Journal of climate* 22.10 (2009), pp. 2501–2511.

[EK06] Martin Erwig and Steve Kollmansberger. “Functional Pearls: Probabilistic functional programming in Haskell”. In: *J. Funct. Program.* 16.1 (2006), pp. 21–34. DOI: 10.1017/S0956796805005721. URL: https://doi.org/10.1017/S0956796805005721.

[Fla11] Gregory M Flato. “Earth system models: an overview”. In: *Wiley Interdisciplinary Reviews: Climate Change* 2.6 (2011), pp. 783–800.

[FP09] Christophe Feltus and Michaël Petit. “Building a responsibility model using modal logic-towards Accountability, Aapability and Commitment concepts”. In: *2009 IEEE/ACS International Conference on Computer Systems and Applications*. IEEE. 2009, pp. 386–391.

[Fra79] Klaus Fraedrich. “Catastrophes and resilience of a zero-dimensional climate system with ice-albedo and greenhouse feedback”. In: *Quarterly Journal of the Royal Meteorological Society* 105.443 (1979), pp. 147–167.

[Ghi14] Michael Ghil. “A Mathematical Theory of Climate Sensitivity or, How to Deal With Both Anthropogenic Forcing and Natural Variability?” In: *Climate Change: Multidecadal and Beyond*. Ed. by C. P. Chang et al. 2014, unknown.

[Ghi76] Michael Ghil. “Climate Stability for a Sellers-Type Model”. In: *Journal of Atmospheric Sciences* 33.1 (1976), pp. 3–20. DOI: 10.1175/1520-0469(1976)033<0003:CSFAST>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/atsc/33/1/1520-0469_1976_033_0003_csfast_2_0_co_2.xml.

[Gib11] J.C. Gibbings. *Dimensional Analysis*. Springer London, 2011. ISBN: 9781849963176. URL: https://books.google.de/books?id=Q6iflrgVaWcC.

[Gin+18] Kees van Ginkel et al. “D3. 1 Operationalizing socio-economic and climate tipping points”. In: *Deliverable of the H2020 COACCH project* (2018).

[Gir81] M. Giry. “A categorial approach to probability theory”. In: *Categorical Aspects of Topology and Analysis*. Ed. by B. Banaschewski. Vol. 915. Lecture Notes in Mathematics. Berlin: Springer, 1981, pp. 68–85.

[GL20] Michael Ghil and Valerio Lucarini. “The physics of climate variability and climate change”. In: *Reviews of Modern Physics* 92.3 (2020), p. 035002.

[Gla00] Malcolm Gladwell. *The tipping point: How little things can make a big difference*. Little, Brown, 2000.

[GM12] Marco Giunti and Claudio Mazzola. “Dynamical systems on monoids: Toward a general theory of deterministic systems and motion”. In: *Methods, models, simulations and approaches towards a general theory of change*. World Scientific, 2012, pp. 173–185.

[Goo+10] Hugues Goosse et al. *Introduction to climate dynamics and climate modeling*. Centre de recherche sur la Terre et le climat Georges Lemaître-UCLouvain, 2010.

[Goo15] Hugues Goosse. *Climate system dynamics and modeling*. Cambridge University Press, 2015.

[Gre+04] J. M. Gregory et al. “A new method for diagnosing radiative forcing and climate sensitivity”. In: *Geophysical Research Letters* 31.3 (2004). DOI: 10.1029/2003GL018747. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2003GL018747. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2003GL018747.

[Han+02] J Hansen et al. “Climate forcings in Goddard Institute for space studies SI2000 simulations”. In: *Journal of Geophysical Research: Atmospheres* 107.D18 (2002), ACL–2.

[Han+05] James Hansen et al. “Efficacy of climate forcings”. In: *Journal of Geophysical Research: Atmospheres* 110.D18 (2005).

[Han+84] J Hansen et al. “Climate sensitivity: Analysis of feedback mechanisms.” In: *feedback* 1 (1984), pp. 1–3.

[Han+85] James Hansen et al. “Climate response times: Dependence on climate sensitivity and ocean mixing”. In: *Science* 229.4716 (1985), pp. 857–859.

[Hei+16] Jobst Heitzig et al. “Topology of sustainable management of dynamical systems with desirable states: from defining planetary boundaries to safe operating spaces in the Earth system”. In: *Earth System Dynamics* 7.1 (2016), pp. 21–50.

[Hei19] Jobst Heitzig. “Efficient non-cooperative provision of costly positive externalities via conditional commitments”. In: *Available at SSRN 3449004* (2019).

[Hei88] Hartmut Heinrich. “Origin and consequences of cyclic ice rafting in the northeast Atlantic Ocean during the past 130,000 years”. In: *Quaternary research* 29.2 (1988), pp. 142–152.

[Hel+10] Isaac M Held et al. “Probing the fast and slow components of global warming by returning abruptly to preindustrial forcing”. In: *Journal of Climate* 23.9 (2010), pp. 2418–2427.

[Hey+16] Anna S von der Heydt et al. “Lessons on climate sensitivity from past climate changes”. In: *Current Climate Change Reports* 2.4 (2016), pp. 148–158.

[HF20] Lukas Halekotte and Ulrike Feudel. “Minimal fatal shocks in multistable complex networks”. In: *Scientific reports* 10.1 (2020), pp. 1–13.

[HK04] Hermann Held and Thomas Kleinen. “Detection of climate system bifurcations by degenerate fingerprinting”. In: *Geophysical Research Letters* 31.23 (2004).

[Ion09] Cezar Ionescu. “Vulnerability Modelling and Monadic Dynamical Systems”. PhD thesis. Freie Universität Berlin, 2009. URL: https://d-nb.info/1023491036/34.

[Ion16] Cezar Ionescu. “Vulnerability modelling with functional programming and dependent types”. In: *Mathematical Structures in Computer Science* 26.1 (2016), pp. 114–128. DOI: 10.1017/S0960129514000139.

[IPC18] IPCC. “Annex I: Glossary in Global Warming of 1.5C. An IPCC Special Report on the impacts of global warming of 1.5C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the thre”. In: (2018).

[Jac15] Bart Jacobs. “New directions in categorical logic, for classical, probabilistic and quantum logic”. In: *Logical Methods in Computer Science* 11 (2015).

[Kéf+13] Sonia Kéfi et al. “Early warning signals also precede non-catastrophic transitions”. In: *Oikos* 122.5 (2013), pp. 641–648.

[KHP03] Thomas Kleinen, Hermann Held, and Gerhard Petschel-Held. “The potential role of spectral properties in detecting thresholds in the Earth system: application to the thermohaline circulation”. In: *Ocean Dynamics* 53.2 (2003), pp. 53–63.

[KLN22] Christian Kuehn, Kerstin Lux, and Alexandra Neamţu. “Warning signs for non-Markovian bifurcations: colour blindness and scaling laws”. In: *Proceedings of the Royal Society A* 478.2259 (2022), p. 20210740.

[Klo+21] Ann Kristin Klose et al. “What do we mean,‘tipping cascade’?” In: *Environmental Research Letters* 16.12 (2021), p. 125011.

[KR11] Peter E Kloeden and Martin Rasmussen. *Nonautonomous dynamical systems.* 176. American Mathematical Soc., 2011.

[KR15] Reto Knutti and Maria AA Rugenstein. “Feedbacks, climate sensitivity and the limits of linear models”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 373.2054 (2015), p. 20150146.

[KRH17] Reto Knutti, Maria AA Rugenstein, and Gabriele C Hegerl. "Beyond equilibrium climate sensitivity". In: *Nature Geoscience* 10.10 (2017), pp. 727–736.

[Kri+13] Gerhard Krinner et al. "Long-term climate change: Projections, commitments and irreversibility". In: *Climate Change 2013 the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* 9781107057999 (2013), pp. 1029–1136. DOI: 10.1017/CBO9781107415324.024.

[Krö+20] Jonathan Krönke et al. "Dynamics of tipping cascades on complex networks". In: *Physical Review E* 101.4 (2020), p. 042311.

[KS02] Reto Knutti and Thomas F Stocker. "Limited predictability of the future thermohaline circulation close to an instability threshold". In: *Journal of Climate* 15.2 (2002), pp. 179–186.

[Kue11] Christian Kuehn. "A mathematical framework for critical transitions: Bifurcations, fast–slow systems and stochastic dynamics". In: *Physica D: Nonlinear Phenomena* 240.12 (2011), pp. 1020–1035.

[Kue13] Christian Kuehn. "A mathematical framework for critical transitions: normal forms, variance and applications". In: *Journal of Nonlinear Science* 23.3 (2013), pp. 457–510.

[Kue15] Christian Kuehn. *Multiple time scale dynamics.* Vol. 191. Springer, 2015.

[Kuz13] Yuri A Kuznetsov. *Elements of applied bifurcation theory.* Vol. 112. Springer Science & Business Media, 2013.

[LD19] Johannes Lohmann and Peter D Ditlevsen. "A consistent statistical model selection for abrupt glacial climate changes". In: *Climate dynamics* 52.11 (2019), pp. 6411–6426.

[LDL12] VN Livina, PD Ditlevsen, and TM Lenton. "An independent test of methods of detecting system states and bifurcations in time-series data". In: *Physica A: Statistical Mechanics and its Applications* 391.3 (2012), pp. 485–496.

[Len+08] Timothy M Lenton et al. "Tipping elements in the Earth's climate system". In: *Proceedings of the national Academy of Sciences* 105.6 (2008), pp. 1786–1793.

[Len+12] TM Lenton et al. "Early warning of climate tipping points from critical slowing down: comparing methods to improve robustness". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 370.1962 (2012), pp. 1185–1204.

[Len+19] Timothy Lenton et al. "Climate tipping points - too risky to bet against". In: *Nature* 575 (2019), pp. 592–595.

[Len12] Timothy M Lenton. "Arctic climate tipping points". In: *Ambio* 41.1 (2012), pp. 10–22.

[Len13] Timothy M Lenton. "Environmental tipping points". In: *Annual Review of Environment and Resources* 38 (2013), pp. 1–29.

[Len20] Timothy M Lenton. "Tipping positive change". In: *Philosophical Transactions of the Royal Society B* 375.1794 (2020), p. 20190123.

[LKL10] Valerie N Livina, F Kwasniok, and Timothy M Lenton. "Potential analysis reveals changing number of climate states during the last 60 kyr". In: *Climate of the Past* 6.1 (2010), pp. 77–82.

[LL07] Valerie N Livina and Timothy M Lenton. "A modified method for detecting incipient bifurcations in a dynamical system". In: *Geophysical research letters* 34.3 (2007).

[Loh+21] Johannes Lohmann et al. "Abrupt climate change as a rate-dependent cascading tipping point". In: *Earth System Dynamics* 12.3 (2021), pp. 819–835.

[Lor63] Edward N Lorenz. "Deterministic nonperiodic flow". In: *Journal of atmospheric sciences* 20.2 (1963), pp. 130–141.

[Lov72] J.E. Lovelock. "Gaia as seen through the atmosphere". In: *Atmospheric Environment (1967)* 6.8 (1972), pp. 579–580. ISSN: 0004-6981. DOI: https://doi.org/10.1016/0004-6981(72)90076-5. URL: https://www.sciencedirect.com/science/article/pii/0004698172900765.

[LS07] Timothy M Lenton and Hans Joachim Schellnhuber. "Tipping the scales". In: *Nature Climate Change* 1.712 (2007), pp. 97–98.

[Ma+22] Jinzhong Ma et al. "Early warning of noise-induced catastrophic high-amplitude oscillations in an airfoil model". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 32.3 (2022), p. 033119.

[Mac78] Saunders MacLane. *Categories for the Working Mathematician*. 2nd. Graduate Texts in Mathematics. Springer, 1978.

[Mar+14] Shaun A Marcott et al. "Centennial-scale changes in the global carbon cycle during the last deglaciation". In: *Nature* 514.7524 (2014), pp. 616–619.

[Mar00] Jochem Marotzke. "Abrupt climate change and thermohaline circulation: Mechanisms and predictability". In: *Proceedings of the National Academy of Sciences* 97.4 (2000), pp. 1347–1350.

[Mar94] Jochem Marotzke. "Ocean models in climate problems". In: *Ocean processes in climate dynamics: Global and mediterranean examples*. Springer, 1994, pp. 79–109.

[Mas+10] Michael D Mastrandrea et al. "Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties". In: (2010).

[Mas+21] V. Masson-Delmotte et al., eds. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (in press)*. Cambridge University Press, 2021.

[Mat+21] J.B.R. Matthews et al. "IPCC 2021: Annex VII: Glossary (in press)". In: *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (in press)* (2021).

[McK+] David I Armstrong McKay et al. "Updated assessment suggests¿ 1.5 C global warming could trigger multiple climate tipping points". In: ().

[Mee+07] Gerald A Meehl et al. *Global climate projections. Climate change 2007: the physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. 2007.

[Mem04] North GRIP Members. "High resolution Climate Record of the Northern Hemisphere reaching into the last Glacial Interglacial Period". In: *Nature* 431 (2004), pp. 147–151.

[Mer] Merriam Webster. *Definition of "System"*. Springfield, MA, USA. http://www.merriam-webster.com/dictionary/system. Retrieved 2019-01-16.

[Mos06] C Moss. *Earth system science in the Anthropocene: emerging issues and problems*. Vol. 103. Springer Science & Business Media, 2006.

[MW75] Syukuro Manabe and Richard T Wetherald. "The effects of doubling the CO2 concentration on the climate of a general circulation model". In: *Journal of Atmospheric Sciences* 32.1 (1975), pp. 3–15.

[MW91] Jochem Marotzke and Jürgen Willebrand. "Multiple equilibria of the global thermohaline circulation". In: *Journal of physical oceanography* 21.9 (1991), pp. 1372–1385.

[Myh+98] Gunnar Myhre et al. "New estimates of radiative forcing due to well mixed greenhouse gases". In: *Geophysical research letters* 25.14 (1998), pp. 2715–2718.

[Oes+84] Hans Oeschger et al. "Late glacial climate history from ice cores". In: *Climate processes and climate sensitivity* 29 (1984), pp. 299–306.

[Pat+18] Frank Pattyn et al. "The Greenland and Antarctic ice sheets under 1.5 °C global warming". In: *Nature Climate Change* 8.12 (2018), pp. 1053–1061. ISSN: 17586798. DOI: 10.1038/s41558-018-0305-8. URL: http://dx.doi.org/10.1038/s41558-018-0305-8.

[Pie03] Roger A Pielke Sr. “Heat storage within the Earth system”. In: *Bulletin of the American Meteorological Society* 84.3 (2003), pp. 331–336.

[Pla+08] Gian Kasper Plattner et al. “Long-term climate commitments projected with climate-carbon cycle models”. In: *Journal of Climate* 21.12 (2008), pp. 2721–2751. ISSN: 08948755. DOI: 10.1175/2007JCLI1905.1.

[Rah96] Stefan Rahmstorf. “On the freshwater forcing and transport of the Atlantic thermohaline circulation”. In: *Climate Dynamics* 12.12 (1996), pp. 799–811.

[Ram88] Veerabhachan Ramanathan. “The greenhouse theory of climate change: a test by an inadvertent global experiment”. In: *Science* 240.4850 (1988), pp. 293–299.

[RN09] Chris Russill and Zoe Nyssa. “The tipping point trend in climate change communication”. In: *Global environmental change* 19.3 (2009), pp. 336–344.

[Roc+09] Johan Rockström et al. “A safe operating space for humanity”. In: *nature* 461.7263 (2009), pp. 472–475.

[Roh+12] Eelco J Rohling et al. “Making sense of palaeoclimate sensitivity”. In: *Nature* 491 (2012), pp. 683–691.

[Roo82] Claes Rooth. “Hydrology and ocean circulation”. In: *Progress in Oceanography* 11.2 (1982), pp. 131–149.

[RS16] Paul Ritchie and Jan Sieber. “Early-warning indicators for rate-induced tipping”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 26.9 (2016), p. 093116.

[Rus11] Chris Russill. “Temporal metaphor in abrupt climate change communication: an initial effort at clarification”. In: *The economic, social and political elements of climate change.* Springer, 2011, pp. 113–132.

[Rus15] Chris Russill. “Climate change tipping points: origins, precursors, and debates”. In: *Wiley Interdisciplinary Reviews: Climate Change* 6.4 (2015), pp. 427–434.

[Ryp16] Martin Rypdal. “Early-warning signals for the onsets of Greenland interstadials and the Younger Dryas–Preboreal transition”. In: *Journal of Climate* 29.11 (2016), pp. 4047–4056.

[Sch+09] Marten Scheffer et al. “Early-warning signals for critical transitions”. In: *Nature* 461.7260 (2009), pp. 53–59.

[Sch83] Michael E Schlesinger. “A review of climate models and their simulation of CO2-induced warming”. In: *International Journal of Environmental Studies* 20.2 (1983), pp. 103–114.

[Sel69] William D Sellers. “A global climatic model based on the energy balance of the earth-atmosphere system”. In: *Journal of Applied Meteorology and Climatology* 8.3 (1969), pp. 392–400.

[Sev+98] Jeffrey P Severinghaus et al. “Timing of abrupt climate change at the end of the Younger Dryas interval from thermally fractionated gases in polar ice”. In: *Nature* 391.6663 (1998), pp. 141–146.

[Sha00] Nicholas J Shackleton. “The 100,000-year ice-age cycle identified and found to lag temperature, carbon dioxide, and orbital eccentricity”. In: *Science* 289.5486 (2000), pp. 1897–1902.

[She+18] Theodore G Shepherd et al. “Storylines: an alternative approach to representing uncertainty in physical aspects of climate change”. In: *Climatic change* 151.3 (2018), pp. 555–571.

[She+20] SC Sherwood et al. “An assessment of Earth’s climate sensitivity using multiple lines of evidence”. In: *Reviews of Geophysics* 58.4 (2020), e2019RG000678.

[SL21] Simon Sharpe and Timothy M Lenton. “Upward-scaling tipping cascades to meet climate goals: Plausible grounds for hope”. In: *Climate Policy* 21.4 (2021), pp. 421–433.

[SM99] Ronald J Stouffer and Syukuro Manabe. “Response of a coupled ocean–atmosphere model to increasing atmospheric carbon dioxide: Sensitivity to the rate of increase”. In: *Journal of Climate* 12.8 (1999), pp. 2224–2237.

[Sol+09] Susan Solomon et al. “Irreversible climate change due to carbon dioxide emissions”. In: *Proceedings of the national academy of sciences* 106.6 (2009), pp. 1704–1709.

[Ste+16] Bjorn Stevens et al. “Prospects for narrowing bounds on Earth’s equilibrium climate sensitivity”. In: *Earth’s Future* 4.11 (2016), pp. 512–522.

[Sto00] Thomas F Stocker. “Past and future reorganizations in the climate system”. In: *Quaternary Science Reviews* 19.1-5 (2000), pp. 301–319.

[Sto04] Ronald J Stouffer. “Time scales of climate response”. In: *Journal of Climate* 17.1 (2004), pp. 209–217.

[Sto11] Thomas Stocker. *Introduction to climate modelling.* Springer Science & Business Media, 2011.

[Sto61] Henry Stommel. “Thermohaline convection with two stable regimes of flow”. In: *Tellus* 13.2 (1961), pp. 224–230.

[Sut81] Alfonso Sutera. “On stochastic perturbation and long-term climate behaviour”. In: *Quarterly Journal of the Royal Meteorological Society* 107.451 (1981), pp. 137–151.

[Trö+17] Jenny Tröltzsch et al. “D1.2 Knowledge synthesis and gap analysis on climate impact analysis, economic costs and scenarios”. In: (2017).

[Tzi00] Eli Tziperman. “Proximity of the present-day thermohaline circulation to an instability threshold”. In: *Journal of Physical Oceanography* 30.1 (2000), pp. 90–104.

[Val11] Paul Valdes. “Built for stability”. In: *Nature Geoscience* 4.7 (2011), pp. 414–416.

[VHS18] Sandra Van der Hel, Iina Hellsten, and Gerard Steen. “Tipping points and climate change: Metaphor between science and the media”. In: *Environmental Communication* 12.5 (2018), pp. 605–620.

[WFR09] EW Wolff, Hubertus Fischer, and R Röthlisberger. “Glacial terminations as southern warmings without northern control”. In: *Nature Geoscience* 2.3 (2009), pp. 206–209.

[Wie+11] S. Wieczorek et al. “Excitability in ramped systems: the compost-bomb instability”. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 467.2129 (2011), pp. 1243–1269. DOI: 10.1098/rspa.2010.0485.

[Wig05] T. M. L. Wigley. “The Climate Change Commitment”. In: *Science* 307.5716 (2005), pp. 1766–1769. ISSN: 0036-8075. DOI: 10.1126/science.1103934. eprint: https://science.sciencemag.org/content/307/5716/1766.full.pdf. URL: https://science.sciencemag.org/content/307/5716/1766.

[Wig84] TML Wigley. In: *Climate Monitor 13* 13 (1984), p. 133.

[Win+20] Ricarda Winkelmann et al. “Social tipping processes for sustainability: An analytical framework”. In: *arXiv preprint arXiv:2010.04488* (2020).

[Wis84] C Wissel. “A universal law of the characteristic return time near thresholds”. In: *Oecologia* 65.1 (1984), pp. 101–107.

[WR01] Tom ML Wigley and Sarah CB Raper. “Interpretation of high projections for global-mean warming”. In: *Science* 293.5529 (2001), pp. 451–454.

[WR93] TML Wigley and SCB Raper. “Future changes in global mean temperature and sea level”. In: *Climate and sea level change: observations, projections and implications* 111 (1993), p. 133.

[WSD01] Richard T Wetherald, Ronald J Stouffer, and Keith W Dixon. “Committed warming and its implications for climate change”. In: *Geophysical Research Letters* 28.8 (2001), pp. 1535–1538.

[Wun+21] Nico Wunderling et al. “Interacting tipping elements increase risk of climate domino effects under global warming”. In: *Earth System Dynamics* 12.2 (2021), pp. 601–619.

I IPCC Glossary entries

(The AR6 WGI glossary is not yet officially approved, but we use it here to make sure to include the most up-to-date information.)

I.1 Concerning climate models

What the IPCC AR6 WGI Glossary tells us about climate models:

IPCC Glossary: Climate model

A qualitative or quantitative representation of the climate system based on the physical, chemical and biological properties of its components, their interactions and feedback processes and accounting for some of its known properties. The climate system can be represented by models of varying complexity; that is, for any one component or combination of components a spectrum or hierarchy of models can be identified, differing in such aspects as the number of spatial dimensions, the extent to which physical, chemical or biological processes are explicitly represented, or the level at which empirical parametrisations are involved. There is an evolution towards more complex models with interactive chemistry and biology. Climate models are applied as a research tool to study and simulate the climate and for operational purposes, including monthly, seasonal and interannual climate predictions. See also Chemistry-climate model, Earth system model (ESM), Earth system model of intermediate complexity (EMIC), Energy balance model (EBM), Simple climate model (SCM), Regional climate model (RCM), Dynamic global vegetation model (DGVM), General circulation model (GCM) and Emulators.

IPCC Glossary: Process-based model

Theoretical concepts and computational methods that represent and simulate the behaviour of real-world systems derived from a set of functional components and their interactions with each other and the system environment, through physical and mechanistic processes occurring over time.

IPCC Glossary: Integrated assessment model (IAM)

Models that integrate knowledge from two or more domains into a single framework. They are one of the main tools for undertaking integrated assessments. One class of IAM used in respect of climate change mitigation may include representations of: multiple sectors of the economy, such as energy, land use and land use change; interactions between sectors; the economy as a whole; associated greenhouse gas (GHG) emissions and sinks; and reduced representations of the climate system. This class of model is used to assess linkages between economic, social and technological development and the evolution of the climate system. Another class of IAM additionally includes representations of the costs associated with climate change impacts, but includes less detailed representations of economic systems. These can be used to assess impacts and mitigation in a cost-benefit framework and have been used to estimate the social cost of carbon.

IPCC Glossary: Simple climate model (SCM)

A broad class of lower-dimensional models of the energy balance, radiative transfer, carbon cycle, or a combination of such physical components. SCMs are also suitable for performing emulations of climate-mean variables of Earth system models (ESMs), given that their structural flexibility can capture both the parametric and structural uncertainties across process-oriented ESM responses. They can also be used to test consistency across multiple lines of evidence with regard to climate sensitivity ranges, transient climate responses (TCRs), transient climate response to cumulative emissions (TCREs) and carbon cycle feedbacks. See also Emulators and Earth system model of intermediate complexity (EMIC).

IPCC Glossary: Earth System Model (ESM)

A coupled atmosphere–ocean general circulation model (AOGCM) in which a representation of the carbon cycle is included, allowing for interactive calculation of atmospheric carbon dioxide (CO_2) or compatible emissions. Additional components (e.g., atmospheric chemistry, ice sheets, dynamic vegetation, nitrogen cycle, but also urban or crop models) may be included. See also Earth system model of intermediate complexity (EMIC).

IPCC Glossary: Earth system Model of Intermediate Complexity (EMIC)

Earth system models of intermediate complexity (EMIC) represent climate processes at a lower resolution or in a simpler, more idealised fashion than an Earth system model (ESM).

IPCC Glossary: Energy balance model (EBM)

An energy balance model is a simplified model that analyses the energy budget of the Earth to compute changes in the climate. In its simplest form, there is no explicit spatial dimension and the model then provides an estimate of the changes in globally averaged temperature computed from the changes in radiation. This zero-dimensional energy balance model can be extended to a one-dimensional or two-dimensional model if changes to the energy budget with respect to latitude, or both latitude and longitude, are explicitly considered.

IPCC Glossary: General circulation model (GCM)

A numerical representation of the atmosphere-ocean-sea ice system based on the physical, chemical and biological properties of its components, their interactions and feedback processes. General circulation models are used for weather forecasts, seasonal to decadal prediction, and climate projections. They are the basis of the more complex Earth system models (ESMs). See also Climate model.

IPCC Glossary: Regional climate model (RCM)

A climate model at higher resolution over a limited area. Such models are used in downscaling global climate results over specific regional domains.

IPCC Glossary: Dynamic global vegetation model (DGVM)

A model that simulates vegetation development and dynamics through space and time, as driven by climate and other environmental changes.

IPCC Glossary: Emulation

Reproducing the behaviour of complex, process-based models (namely, Earth System Models, ESMs) via simpler approaches, using either emulators or simple climate models (SCMs). The computational efficiency of emulating approaches opens new analytical possibilities given that ESMs take a lot of computational resources for each simulation. See also Emulators and Simple climate model (SCM).

IPCC Glossary: Emulators

A broad class of heavily parametrized models (‘one-or-few-line climate models’), statistical methods like neural networks, genetic algorithms or other artificial intelligence approaches, designed to reproduce the responses of more complex, process-based Earth System Models (ESMs). The main application of emulators is to extrapolate insights from ESMs and observational constraints to a larger set of emission scenarios. See also Emulation and Simple climate model (SCM).

I.2 Concerning model experiments

IPCC Glossary: Equilibrium and transient climate experiment

An equilibrium climate experiment is a climate model experiment in which the model is allowed to fully adjust to a change in radiative forcing. Such experiments provide information on the difference between the initial and final states of the model, but not on the time-dependent response. If the forcing is allowed to evolve gradually according to a prescribed emission scenario, the time-dependent response of a climate model may be analysed. Such an experiment is called a transient climate experiment. See also *Climate projection*. (AR5, WGI)

IPCC Glossary: Model initialization

A climate prediction typically proceeds by integrating a climate model forward in time from an initial state that is intended to reflect the actual state of the climate system. Available observations of the climate system are 'assimilated' into the model. Initialization is a complex process that is limited by available observations, observational errors and, depending on the procedure used, may be affected by uncertainty in the history of climate forcing. The initial conditions will contain errors that grow as the forecast progresses, thereby limiting the time for which the forecast will be useful.

IPCC Glossary: Parameterisation

In climate models, this term refers to the technique of representing processes that cannot be explicitly resolved at the spatial or temporal resolution of the model (sub-grid scale processes) by relationships between model-resolved larger-scale variables and the area- or time-averaged effect of such subgrid scale processes.

I.3 Concerning model simulations/scenarios/pathways

IPCC Glossary: Pathways

The temporal evolution of natural and/or human systems towards a future state. Pathway concepts range from sets of quantitative and qualitative scenarios or narratives of potential futures to solution-oriented decision-making processes to achieve desirable societal goals. Pathway approaches typically focus on biophysical, techno-economic, and/or socio-behavioural trajectories and involve various dynamics, goals, and actors across different scales. See also Scenario storyline (under Storyline), Mitigation scenario (under Scenario), Baseline scenario (under Scenario) and Stabilisation (of GHG or CO_2-equivalent concentration).

$1.5^\circ\mathrm{C}$*pathway* A pathway of emissions of greenhouse gases and other climate forcers that provides an approximately one-in-two to two-in-three chance, given current knowledge of the climate response, of global warming either remaining below $1.5^\circ\mathrm{C}$ or returning to $1.5^\circ\mathrm{C}$ by around 2100 following an overshoot.

Representative concentration pathways (RCPs) Scenarios that include time series of emissions and concentrations of the full suite of greenhouse gases (GHGs) and aerosols and chemically active gases, as well as land use/land cover (Moss et al., 2010). The word representative signifies that each RCP provides only one of many possible scenarios that would lead to the specific radiative forcing characteristics. The term pathway emphasises that not only the long-term concentration levels are of interest, but also the trajectory taken over time to reach that outcome (Moss et al., 2010).

RCPs usually refer to the portion of the concentration pathway extending up to 2100, for which Integrated assessment models produced corresponding emission scenarios. Extended concentration pathways describe extensions of the RCPs from 2100 to 2300 that were calculated using simple rules generated by stakeholder consultations, and do not represent fully consistent scenarios. Four RCPs produced from Integrated assessment models were selected from the published literature and are used in the Fifth IPCC Assessment and also used in this Assessment for comparison, spanning the range from approximately below $2°\mathrm{C}$ warming to high ($> 4°\mathrm{C}$) warming best-estimates by the end of the 21st century: RCP2.6, RCP4.5 and RCP6.0 and RCP8.5.

- *RCP2.6*: One pathway where radiative forcing peaks at approximately $3\mathrm{W\,m^{-2}}$ and then declines to be limited at $2.6\mathrm{W\,m^{-2}}$ in 2100 (the corresponding Extended Concentration Pathway, or ECP, has constant emissions after 2100).
- *RCP4.5* and *RCP6.0*: Two intermediate stabilisation pathways in which radiative forcing is limited at approximately $4.5\mathrm{W\,m^{-2}}$ and $6.0\mathrm{W\,m^{-2}}$ in 2100 (the corresponding ECPs have constant concentrations after 2150).
- *RCP8.5*: One high pathway which leads to $> 8.5\mathrm{W\,m^{-2}}$ in 2100 (the corresponding ECP has constant emissions after 2100 until 2150 and constant concentrations after 2250).

See also Coupled Model Intercomparison Project (CMIP) and Shared socio-economic pathways (SSPs) (under Pathways).
Shared socio-economic pathways (SSPs) Shared socio-economic pathways (SSPs) have been developed to complement the Representative concentration pathways (RCPs). By design, the RCP emission and concentration pathways were stripped of their association with a certain socio-economic development. Different levels of emissions and climate change along the dimension of the RCPs can hence be explored against the backdrop if different socio-economic development pathways (SSPs) on the other dimension in a matrix. This integrative SSP-RCP framework is now widely used in the climate impact and policy analysis literature, where climate projections obtained under the RCP scenarios are analysed against the backdrop of various SSPs. As several emission updates were due, a new set of emission scenarios was developed in conjunction with the SSPs. Hence, the abbreviation SSP is now used for two things: On the one hand SSP1, SSP2, ..., SSP5 is used to denote the five socio-economic scenario families. On the other hand, the abbreviations SSP1-1.9, SSP1-2.6, ..., SSP5-8.5 are used to denote the newly developed emission scenarios that are the result of an SSP implementation within an integrated assessment model. Those SSP scenarios are bare of climate policy assumption, but in combination with so-called shared policy assumptions (SPAs), various approximate radiative forcing levels of 1.9, 2.6, ..., or $8.5\mathrm{W\,m^{-2}}$ are reached by the end of the century, respectively.

IPCC Glossary: Projection

A potential future evolution of a quantity or set of quantities, often computed with the aid of a model. Unlike predictions, projections are conditional on assumptions concerning, for example, future socio-economic and technological developments that may or may not be realised. See also Climate projection, Pathways and Scenario.

IPCC Glossary: Scenario

A plausible description of how the future may develop based on a coherent and internally consistent set of assumptions about key driving forces (e.g., rate of technological change (TC), prices) and relationships. Note that scenarios are neither predictions nor forecasts, but are used to provide a view of the implications of developments and actions. See also Climate scenario and Regional climate scenario.
Baseline scenario See Reference ScenarioSee Reference scenario (under Scenario).
Concentrations scenario A plausible representation of the future development of atmospheric concentrations of substances that are radiatively active (e.g., greenhouse gases (GHGs), aerosols, tropospheric ozone), plus human-induced land cover changes that can be radiatively active via albedo changes, and often used as input to a climate model to compute climate projections.

Emissions scenario A plausible representation of the future development of emissions of substances that are radiatively active (e.g., greenhouse gases (GHGs) or aerosols), plus human-induced land cover changes that can be radiatively active via albedo changes, based on a coherent and internally consistent set of assumptions about driving forces (such as demographic and socio-economic development, technological change, energy and land use) and their key relationships. Concentration scenarios, derived from emission scenarios, are often used as input to a climate model to compute climate projections. See also Representative concentration pathways (RCPs) (under Pathways) and Shared socio-economic pathways (SSPs) (under Pathways).
Mitigation scenario A plausible description of the future that describes how the (studied) system responds to the implementation of mitigation policies and measures. See also Pathways, Socio-economic scenario (under Scenario) and Stabilisation (of GHG or CO_2-equivalent concentration).

Reference scenario Scenario used as starting or reference point for a comparison between two or more scenarios.

Note 1: In many types of climate change research, reference scenarios reflect specific assumptions about patterns of socio-economic development and may represent futures that assume no climate policies or specified climate policies, for example those in place or planned at the time a study is carried out. Reference scenarios may also represent futures with limited or no climate impacts or adaptation, to serve as a point of comparison for futures with impacts and adaptation. These are also referred to as baseline scenarios in the literature.

Note 2: Reference scenarios can also be climate policy or impact scenarios, which in that case are taken as a point of comparison to explore the implications of other features, e.g., of delay, technological options, policy design and strategy or to explore the effects of additional impacts and adaptation beyond those represented in the reference scenario.

Note 3: The term business as usual scenario has been used to describe a scenario that assumes no additional policies beyond those currently in place and that patterns of socio-economic development are consistent with recent trends. The term is now used less frequently than in the past.

Note 4: In climate change attribution or impact attribution research reference scenarios may refer to counterfactual historical scenarios assuming no anthropogenic greenhouse gas emissions (climate change attribution) or no climate change (impact attribution).
Socio-economic scenario A scenario that describes a plausible future in terms of population, gross domestic product (GDP), and other socio-economic factors relevant to understanding the implications of climate change. See also Baseline scenario (under Scenario), Mitigation scenario (under Scenario) and Pathways.

IPCC Glossary: Regional climate scenario

A narrative used to describe how the future might unfold for a region (IPCC-TGICA, 2007). These are often used to guide impact understanding and adaptation efforts. They can include quantitative information based on scaled historical data or derived from GCM-based internally consistent future climates.

See also Climate scenario.

IPCC Glossary: Storyline

A way of making sense of a situation or a series of events through the construction of a set of explanatory elements. Usually it is built on logical or causal reasoning. In climate research, the term storyline is used both in connection to scenarios as related to a future trajectory of the climate and human systems or to a weather or climate event. In this context, storylines can be used to describe plural, conditional possible futures or explanations of a current situation, in contrast to single, definitive futures or explanations.

Physical climate storyline A self-consistent and plausible unfolding of a physical trajectory of the climate system, or a weather or climate event, on timescales from hours to multiple decades (Shepherd et al., 2018). Through this, storylines explore, illustrate and communicate uncertainties in the climate system response to forcing and in internal variability.

Scenario storyline A narrative description of a scenario (or family of scenarios), highlighting the main scenario characteristics, relationships between key driving forces and the dynamics of their evolution.

I.4 Concerning climate sensitivity

IPCC Glossary: Climate metrics

Measures of aspects of the overall climate system response to radiative forcing, such as equilibrium climate sensitivity (ECS), transient climate response (TCR), transient climate response to cumulative CO 2 emissions (TCRE) and the airborne fraction of anthropogenic carbon dioxide. See also Greenhouse gas emission metric, Climate indicator and Key climate indicators (under Climate indicator).

IPCC Glossary: Climate sensitivity

The change in the surface temperature in response to a change in the atmospheric carbon dioxide (CO_2) concentration or other radiative forcing. See also Climate feedback parameter.

IPCC Glossary: Earth system sensitivity

The equilibrium surface temperature response of the coupled atmosphere-ocean- cryosphere-vegetation-carbon cycle system to a doubling of the atmospheric carbon dioxide (CO_2) concentration is referred to as Earth System sensitivity. Because it allows ice sheets to adjust to the external perturbation, it may differ substantially from the equilibrium climate sensitivity derived from coupled atmosphere-ocean models.

IPCC Glossary: Equilibrium climate sensitivity (ECS)

The equilibrium (steady state) change in the surface temperature following a doubling of the atmospheric carbon dioxide (CO_2) concentration from pre-industrial conditions.

Closely related:

IPCC Glossary: Climate feedback parameter

A way to quantify the radiative response of the climate system to a global surface temperature change induced by a radiative forcing. It is quantified as the change in net energy flux at the top of atmosphere for a given change in annual global surface temperature. It has units of $\mathrm{W\,m^{-2}\,{}^{\circ}C^{-1}}$.

I.5 Concerning climate commitment

IPCC Glossary: Climate change commitment

Climate change commitment is defined as the unavoidable future climate change resulting from inertia in the geophysical and socio-economic systems. Different types of climate change commitment are discussed in the literature (see subterms). Climate change commitment is usually quantified in terms of the further change in temperature, but it includes other future changes, for example in the hydrological cycle, in extreme weather events, in extreme climate events, and in sea level.

Constant composition commitment The constant composition commitment is the remaining climate change that would result if atmospheric composition, and hence radiative forcing, were held fixed at a given value. It results from the thermal inertia of the ocean and slow processes in the cryosphere and land surface.

Constant emissions commitment The constant emissions commitment is the committed climate change that would result from keeping anthropogenic emissions constant.

Zero emissions commitment The zero emissions commitment is an estimate of the subsequent global warming that would result after anthropogenic emissions are set to zero. It is determined by both inertia in physical climate system components (ocean, cryosphere, land surface) and carbon cycle inertia. In its widest sense it refers to emissions of each climate forcer including greenhouses gases, aerosols and their pre- cursors. The climate response to this can be complex due to the different timescale of response of each climate forcer. A specific sub-category of zero emissions commitment is the Zero CO_2 Emissions Commitment which refers to the climate system response to CO_2 emissions after setting these to net zero. The CO_2-only definition is of specific use in estimating remaining carbon budgets.

I.6 Concerning abrupt climate change

IPCC Glossary: Abrupt change

A change in the system that is substantially faster than the typical rate of the changes in its history.

IPCC Glossary: Abrupt climate change

A large-scale abrupt change in the climate system that takes place over a few decades or less, persists (or is anticipated to persist) for at least a few decades and causes substantial impacts in human and/or natural systems.

IPCC Glossary: Tipping element

A component of the Earth System that is susceptible to a tipping point.

IPCC Glossary: Tipping point

A critical threshold beyond which a system reorganizes, often abruptly and/or irreversibly.

IPCC Glossary: Irreversibility

A perturbed state of a dynamical system is defined as irreversible on a given timescale, if the recovery from this state due to natural processes takes substantially longer than the timescale of interest.

II TiPES work package objectives

In this appendix we list the objectives of WP1–5 for reference.

WP1: Observation-based analysis of tipping elements

1. To provide the empirical basis to study abrupt climatic transitions that have occurred in past warm and cold climates, focusing on proxy data synthesis and synchronization of different records.
2. To derive probabilistic time series representations of proxy records that allow for a mathematically rigorous propagation of dating uncertainties to subsequent analysis such as synchronization and dependency analyses between different records, search for EWS, but also the model evaluations planned in WP2.
3. To assess interactions between different TEs, and the ecological and societal impacts of past abrupt climate transitions. This will provide valuable information for estimating the impacts of potentially similar events due to global warming in the future.
4. To extend existing concepts of statistical EWSs in paleoclimatic proxy records by advancing the employed statistical estimators and taking into account associated physical mechanisms. This will provide detailed information regarding the specific subsystems and statistical characteristics in which EWSs for potential future abrupt transitions should be searched for.

WP2: Modelling of tipping elements in past climate

1. To assess the climate stability of a suite of state-of-the-art climate models. These are the EMICs Bern3D, CLIMBER-X, and FAMOUS, as well as the ESMs CESM, HadCM3, and UKESM1. This goal primarily feeds into TiPES Objective 1.
2. In delivering this assessment on model stability, to ensure that all WP2 work has the maximum positive impact on, and contribution to, the IPCC process.
3. To implement proxies directly in models to ensure that they can be evaluated against the WP1 datasets. This will help maintain European leadership in the research area of ESM isotope code development.

WP3: Modelling of tipping elements in present and future climate

1. To assess the risk of an AMOC shutdown in the near future due to global warming, and to identify corresponding EWSs from simulations.
2. To quantify the risk that the Amazon rainforest will tip to a savannah state due to global-warming-induced precipitation changes and due to deforestation, to identify EWSs for such a transition, and to provide an assessment of the climatological, ecological and socio-economic impacts of a potential dieback of the Amazon.

3. To estimate the response of the Indian summer monsoon to various scenarios of regional and global forcings in order to allow to define safe operating spaces that exclude future dangerous conditions in the hydroclimatology of the region, and to quantify EWSs of a potential regime shift of the monsoon.
4. To assess the impacts of TP crossings on a range of relevant climate properties in Europe, such as the statistics of atmospheric blocking, weather regimes and extremes, to identify worst case scenarios in this regard, and to establish associated EWS.
5. To constrain the stability and future ice-sheet evolution in the mid-long term of the Antarctic and Greenland ice sheets, and to identify suitable precursor signals of a future stability loss.

WP4: Climate sensitivity and response

1. To extend the concept of climate sensitivity in order to systematically allow for state-dependent feedbacks, a large set of climatic observables, and multiple spatial and temporal scales.
2. To incorporate the effect of TPs in defining climate sensitivity.

WP5: Theoretical underpinning of tipping points

1. To understand the robustness of TPs across the climate model hierarchy.
2. To further develop and apply non-autonomous dynamical systems theory appropriate to understand climate tipping phenomena in the presence of a variety of GHG emissions scenarios.
3. To develop and evaluate novel early warning signals (EWS) for non-autonomous climate problems and associated statistical forecasting and detection tests.

A DSL for Monadic Decision Problems, Responsibility under Uncertainty and Tipping Point Notions

Nuria Brede[1,2], Nicola Botta[1,3], Michel Crucifix[4], and Marina Martínez Montero[4]

[1]RD4: Complexity Science, Potsdam Institute for Climate Impact Research, Potsdam, Germany
[2]Department of Computer Science, University of Potsdam, Potsdam, Germany
[3]Department of Computer Science and Engineering, Chalmers University of Technology, Göteborg, Sweden
[4]Earth and Life Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium

{nuria.brede,botta}@pik-potsdam.de, {michel.crucifix,marina.martinez}@uclouvain.be

Abstract

We develop a domain-specific language (DSL) for the specification of decision problems in the context of tipping point research, on top of a lightweight version the generic Botta et al. 2017 framework for specifying and solving monadic sequential decision problems. The aim is to improve accountability in the context of climate policy advice by narrowing the gap between mathematical problem specification and implementation. This is achieved by using a programming language based on Dependent Type Theory, in which it is possible to express specification, implementation and proof that the implementation fulfils certain properties within the same language.

We extend the Botta et al. theory with generic measures of responsibility and a syntax to transparently express goals of decision making. The usage of the framework is illustrated by the specification of a conceptual stochastic greenhouse gas emission problem. In a further extension of the basic theory, we show the correctness of the generic backward induction algorithm implemented in the framework in a more general setting than commonly considered in control theory.

Contents

1 Introduction

We develop a domain-specific language (DSL) for the specification of decision problems in the context of tipping point research based on earlier work on a computational theory of policy advice proposed by Botta et al. in [BJI17a]. Our aim is to improve accountability in the context of climate policy advice by narrowing the gap between mathematical problem specification and implementation [BBCMM22b, BBCMM21, BBCMM22a] This is achieved by using a programming language based on Dependent Type Theory, in which it is possible to express specification, implementation and proof that the implementation fulfils certain properties within the same language.

As basis for the further work presented in this report, we define a lightweight version of the [BJI17a] theory that is easier to use in practice and yet sufficiently expressive. We extend the Botta et al. theory with generic measures of responsibility and a syntax to transparently express goals of decision making. The usage of the framework is illustrated by the specification of a conceptual stochastic climate policy problem. In the example we also discuss how to make use of conditional probabilities in the spirit of Bayesian belief networks to define transition function of stochastic decision problems in a modular way.

In a further extension of the basic theory, we discuss what correctness means for the generic backward induction algorithm of the theory and show under which conditions it computes provably optimal policy sequences. This allows to use this algorithm in a more general setting than commonly considered in control theory. This is crucial if non-standard combinations of measures and non-determinism, beyond the expected value measure and ordinary stochastic problems are considered.

In this report we primarily summarise our work presented in [BB21, BBC$^+$21] including some small extensions, and give pointers to possible future work. Together with [BBCMM22c] it forms Deliverable 6.2 of the EU Horizon 2020 Project TiPES [TiP23]. As a supplement, the slides of several introductory talks on the material covered in this report and the underlying paradigm are available online [Bot20a, Bot20b, Bot20c, Bre20, MMB20, Bot21, Bre21, Bot22, Bre22].

Technical remarks. The theory used in the report is heavily based on dependent types and is formulated in the programming language *Idris* [Bra17, The10]. Public versions in Idris are available in [B$^+$21] and [B$^+$22]. For introductions to functional programming and dependent types, see [Bir14, Bra17]. An introductory course on formal specification, monadic dynamical systems and the IdrisLibs framework of [BJI17a] is available as TiPES deliverable D6.1 [BBCMM20]. The report itself has been generated via lhs2TEX[HL15] from literate Idris files. These are publicly available at `https://doi.org/10.5281/zenodo.6826927` and can be type-checked for correctness.

2 Framework

In this section, we give a summary of the theory for the specifying and solving of *finite horizon sequential decision problems* (*SDP*s) which is used in [BB21] and [BBC$^+$21]. This theory is a lightweight version of the framework of Botta, Jansson and Ionescu presented in [BJI17a] which has been developed in TiPES, trading to a certain degree expressivity against ease of use. Here we recap the lightweight theory and briefly discuss the differences between the two theories in Subsection 2.2. Longer introductions to monadic sequential decision problems and the respective versions of the framework can be found in [BJI17a, BB21, BBC$^+$21].

2.1 Problem specification and solution

In a nutshell, the theory consists of two sets of components: one for the *specification* of sequential decision problems (SDPs) and one for their *solution* with verified backward induction. For informal introductions to SDPs, see [BJI17a]. Reference mathematical introductions to SDP are given in sections 1.2 and 2.1 of [Ber95] and [Put14], respectively.

Specification components. The specification of an SDP comprises three parts. The first part consists of four components that specify the sequential decision *process* that underlies a decision *problem*:

- A *monad* M, accounting for the uncertainties that affect the decision process:[1].

  $M : \mathit{Type} \rightarrow \mathit{Type}$

- A type family X

  $X : (t : \mathbb{N}) \rightarrow \mathit{Type}$

  where $X\ t$ is the type of states at decision step t

- A type family Y

  $Y : (t : \mathbb{N}) \rightarrow X\ t \rightarrow \mathit{Type}$

  where $Y\ t\ x$ is the type of controls available at decision step t and state x

- A transition function

  $\mathit{next} : (t : \mathbb{N}) \rightarrow (x : X\ t) \rightarrow Y\ t\ x \rightarrow M\ (X\ (S\ t))$

  such that $\mathit{next}\ t\ x\ y$ is an M-structure of the states that can be reached by selecting control y in state x at decision step t.

The uncertainty monad, the states, the controls and the next function completely specify a decision process: if we were given a rule for selecting controls for a given decision process (that is, a function that gives us a control for every possible state) and an initial state (or a probability distribution of initial states) we could, in principle, compute all possible trajectories compatible with that initial state (or with that probability distribution) together with their probabilities.

Indeed, a sequential decision problem for n steps consists of finding a sequence of n *policies* (in control theory, functions that map states to controls or, in other words, decision rules, are called policies) that, for a given decision process, maximises the value of taking n decision steps according to those policies, one after the other.

In turn, the value of taking n decision steps according to a sequence of n policies is defined through a measure (in stochastic problems often the expected-value measure) of a sum of rewards obtained along the trajectories.

It follows that, in order to fully specify a decision problem, one has to define the rewards obtained at each decision step, the sum that the decision maker seeks to maximise and the measure function. This is done in terms of six problem specification components that form the second part of the specification.

- A type of values

  $\mathit{Val} : \mathit{Type}$

[1] Discussing the notion of monad here would go beyond the scope of this report, but see [Wad92] for an introduction to monads in computer science, and our papers [BBJR21] and [BB21, Appendix 1.2] for their application in the context of the Botta et al. framework.

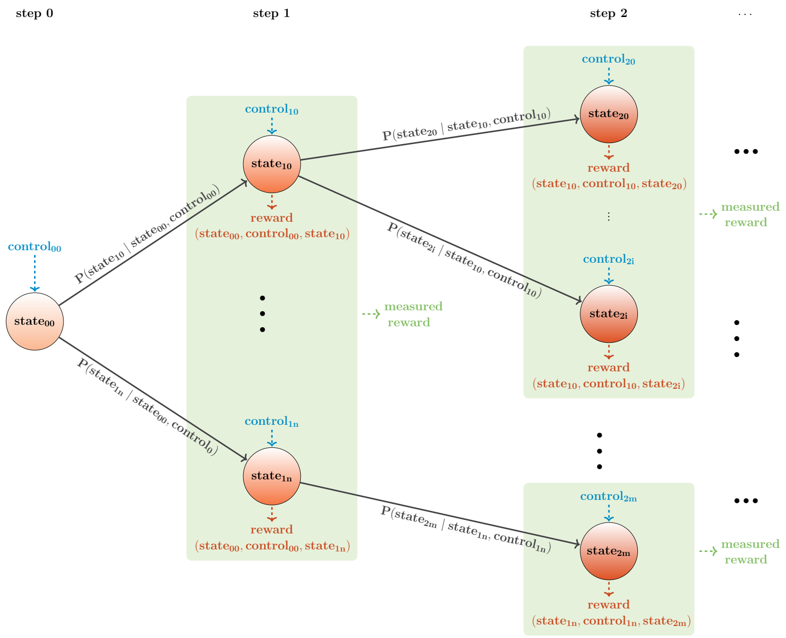

step 0
step 1
step 2
...
control$_{00}$
state$_{00}$
P(state$_{10}$ | state$_{00}$, control$_{00}$)
P(state$_{1n}$ | state$_{00}$, control$_{0}$)
control$_{10}$
state$_{10}$
reward
(state$_{00}$, control$_{00}$, state$_{10}$)
control$_{1n}$
state$_{1n}$
reward
(state$_{00}$, control$_{00}$, state$_{1n}$)
measured
reward
P(state$_{20}$ | state$_{10}$, control$_{10}$)
P(state$_{2i}$ | state$_{10}$, control$_{10}$)
P(state$_{2m}$ | state$_{1n}$, control$_{1n}$)
control$_{20}$
state$_{20}$
reward
(state$_{10}$, control$_{10}$, state$_{20}$)
control$_{2i}$
state$_{2i}$
reward
(state$_{10}$, control$_{10}$, state$_{2i}$)
control$_{2m}$
state$_{2m}$
reward
(state$_{1n}$, control$_{1n}$, state$_{2m}$)
measured
reward
measured
reward

Figure 1: Schematic illustration of a stochastic SDP.

- A reward function

$$\mathit{reward} : (t : \mathbb{N}) \to (x : X\ t) \to Y\ t\ x \to X\ (S\ t) \to \mathit{Val}$$

  $\mathit{reward}\ t\ x\ y\ x'$ is the reward obtained by selecting control y in state x when the next state is x'

- A binary operation

$$(\oplus) : \mathit{Val} \to \mathit{Val} \to \mathit{Val}$$

  for aggregating rewards

- a reference value

$$\mathit{zero} : \mathit{Val}$$

  determining an initial reward (or cost) before taking any decision[2]

- A measure

$$\mathit{meas} : M\ \mathit{Val} \to \mathit{Val}$$

- A total preorder

$$(\sqsubseteq) : \mathit{Val} \to \mathit{Val} \to \mathit{Type}$$

  that allows to compare values.

A few remarks are at place here.

1. In many applications, *Val* is a numerical type and controls represent amounts of used resources like fuel, water, etc. In these cases, the reward function encodes the value (cost) of these resources (and perhaps also the benefits achieved by using them) over a decision step. Often, the latter also depend both on the "current" state x and on the next state x'.

[2] The name might suggest that *zero* is supposed to be a neutral element relative to $\oplus$. However, this is not required by the framework.

2. When Val is a numerical type, $\oplus$ is often the canonical addition associated with that type. However, in many applications more flexibility is needed, e.g., to model that decision makers value later rewards less than earlier ones. Again, formulating the theory in terms of a generic addition rule nicely covers all these applications.

3. Mapping *reward t x y* onto *next t x y* (remember that M is a monad and thus a functor) yields a value of type $M\ Val$. These are the *possible* rewards obtained by selecting control y in state x at decision step t.

4. In mathematical theories of optimal control, Val often is $\mathbb{R}$, M is a probability monad and the probability distributions on real numbers are compared based on the *expected value measure*. See Figure 1 for a schematic illustration of a stochastic SDP.

5. In many applications, most prominently in climate policy, measuring uncertainty of rewards in terms of expected value measures is inadequate. This is why the theory provides the possibility to use other measures (and monads) as well. However, combinations of measure, monad and $\oplus$ need to fulfil certain compatibility conditions which will be discussed in Section 4.

The third part of a specification concerns the axioms that the data components have to fulfil. These parts are needed for the verification of the monadic backward induction algorithm which will be discussed in Section 4.

- Monad structure of M

$$monadM : Monad\ M$$

- The relation $\sqsubseteq$ is a total preorder

$$lteTP : TotalPreorder\ (\sqsubseteq)$$

- Monotonicity axioms for $\oplus$ and *meas*

$$\begin{aligned} plusMon\ &: \{v1, v2, v3, v4 : Val\} \to \\ &\quad v1 \sqsubseteq v2 \to v3 \sqsubseteq v4 \to (v1 \oplus v3) \sqsubseteq (v2 \oplus v4) \\ measMon &: Functor\ M \Rightarrow \{A : Type\} \to \\ &\quad (f, g : A \to Val) \to ((a : A) \to f\ a \sqsubseteq g\ a) \to \\ &\quad (ma : M\ A) \to meas\ (map\ f\ ma) \sqsubseteq meas\ (map\ g\ ma) \end{aligned}$$

- Compatibility conditions for M, *meas* and $\oplus$:

  - The measure needs to be left-inverse to *pure*: [3]

  $$measPureSpec : Monad\ M \Rightarrow meas \circ pure \doteq id$$

  - Applying the measure after *join* needs to be extensionally equal to applying it after *map meas*:

  $$measJoinSpec : Monad\ M \Rightarrow meas \circ join \doteq meas \circ map\ meas$$

  - For arbitrary $v : Val$ and non-empty $mv : M\ Val$ applying the measure after mapping $(v\oplus)$ onto mv needs to be equal to applying $(v\oplus)$ after the measure:

  $$\begin{aligned} measPlusSpec : Monad\ M \Rightarrow (v : Val) \to (mv : M\ Val) \to (NonEmpty\ mv) \to \\ (meas \circ map\ (v\oplus))\ mv = ((v\oplus) \circ meas)\ mv \end{aligned}$$

[3] The symbol $\doteq$ denotes *extensional* equality, see [BBJR21] and [BB21, Appendix 1.3].

For convenience, we may package up the above in records or type classes (called *interfaces* in Idris), so that we can work with more than one SDP instance at a time.

```
interface Monad M ⇒ MSDProcess (M : Type → Type)
  (X : (t : ℕ) → Type) (Y : (t : ℕ) → X t → Type) where
  next : (t : ℕ) → (x : X t) → Y t x → M (X (S t))
```

The definition of a monadic SDP requires the value type *Val* to carry some algebraic structure captured by more general interfaces

```
interface Pointed (T : Type) where
  point : T
interface Preorder (T : Type) where
  (⩽) : T → T → Type

interface Monad M ⇒ MAlgebra (M : Type → Type) (T : Type) where
  alg : M T → T
```

The Idris standard library already has an interface for semigroups amounting to

```
interface Semigroup (T : Type) where
  (⊕) : T → T → T
```

Now we can define an interface for monadic SDPs:

```
interface (MSDProcess M X Y,
  Pointed Val, Preorder Val, Semigroup Val, MAlgebra M Val) ⇒
  MSDProblem (M : Type → Type)
  (X : (t : ℕ) → Type) (Y : (t : ℕ) → X t → Type)
  (Val : Type) where
  reward : (t : ℕ) → (x : X t) → Y t x → X (S t) → Val
```

These interfaces only encapsulate the data part of the definitions. For verification purposes we could moreover define interfaces for the axioms that have to be fulfilled. E.g. for functors (prerequisite for monads):

```
interface Functor F ⇒ VeriFunctor (F : Type → Type) where
  mapPresId   : {A : Type} → ExtEq {A = F A} (map id) id
  mapPresComp : {A, B, C : Type} → (f : A → B) → (g : B → C) →
                ExtEq {A = F A} (map (g ∘ f)) (map g ∘ map f)
  mapPresEE   : {A, B : Type} → (f, g : A → B) →
                ExtEq f g → ExtEq {A = F A} (map f) (map g)
```

For the case of functors and monads we have discussed design choices that go into the definition of such verification interfaces in [BBJR21], but in this report we will not go further into this issue.

Solution components. The second set of components of the theory is a generic formalisation of classical optimal control theory for sequential decision problems. Here, we recall the central elements. Motivation for the formalisation can be found in [BJI$^+$17b], [BJI17a] and [BJI18]. For an introduction to the mathematical theory of optimal control, we recommend [Put14] and [Ber95].

The basic notion of control theory is that of a *policy* – a decision rule. Policies are functions from states to controls:

```
Policy : (t : ℕ) → Type
Policy t = (x : X t) → Y t x
```

Policy sequences of length $n : \mathbb{N}$ then are essentially just vectors of policies.[4]

[4] Note that in Idris, S and Z are the constructors of the data type of natural numbers $\mathbb{N}$. Arguments in curly brackets like $\{t : \mathbb{N}\}$ in the definition of *Nil* and (::) are *implicit* parameters. If they can be inferred from the context, they don't have to be given as arguments later on. For (::) this allow to write policy sequences composed of a policy p and as policy sequence *ps* simply as $p :: ps$.

```
data PolicySeq : (t : ℕ) → (n : ℕ) → Type where
  Nil : {t : ℕ} → PolicySeq t Z
  (::) : {t, n : ℕ} → Policy t → PolicySeq (S t) n → PolicySeq t (S n)
```

Perhaps, the most important notion in the mathematical theory of optimal control is that of *value function*. The value function takes two arguments: a policy sequence *ps* for making n decision steps starting from decision step t and an initial state in $x : X\ t$. It computes the value of taking n decision steps according to the policies *ps* when starting in x:

```
val : Functor M ⇒ {t, n : ℕ} → PolicySeq t n → X t → Val
val {t} Nil x       = zero
val {t} (p :: ps) x = let y   = p x in
                      let mx′ = next t x y in
                      meas (map (reward t x y ⨁ val ps) mx′)
```

where

```
(⨁) : {A : Type} → (f, g : A → Val) → A → Val
f ⨁ g = λa ⇒ f a ⊕ g a
```

is a lifted version of $\oplus$. Notice that, independently of the initial state x, the value of the empty policy sequence is *zero*, the problem-specific reference value that has to be provided as part of a problem specification.

The value of a policy sequence consisting of a first policy p and of a tail policy sequence *ps* is defined inductively as the measure of a M-structure of *Val* values. These values are obtained by first computing the control y dictated by p in x, the M-structure of possible next states mx' dictated by *next* and finally by adding *reward* $t\ x\ y\ x'$ and *val ps* x' for all x' in mx'. The result of this functorial mapping is then measured with the problem-specific measure *meas* to obtain a result of type *Val*.

As shown in [BB21], *val ps x* does compute the *meas*-measure of the $\oplus$-sum of the *reward*-rewards along the possible trajectories starting at x under *ps* for sound choices of *meas*. We will come back to this in Section 4.

The above definition of *val* can be exploited to compute policy sequences that are provably optimal. This observation was originally made by Bellman [Bel57] for deterministic and stochastic SDPs. Provided that we can compute optimal extensions of arbitrary policy sequences

```
bestExt : Functor M ⇒ {t, n : ℕ} → PolicySeq (S t) n → Policy t
```

it is easy to derive a generic implementation of backward induction:[5]

```
bi : Functor M ⇒ (t : ℕ) → (n : ℕ) → PolicySeq t n
bi t Z     = Nil
bi t (S n) = let ps = bi (S t) n in bestExt ps :: ps
```

This implementation of backward induction can be proven to compute *optimal policy sequences* if *bestExt* computes *optimal extensions* of policy sequences:

```
BestExt : Functor M ⇒ {t, n : ℕ} → PolicySeq (S t) n → Policy t → Type
BestExt {t} ps p = (p′ : Policy t) → (x : X t) → val (p′ :: ps) x ⊑ val (p :: ps) x


bestExtSpec : Functor M ⇒ {t, n : ℕ} → (ps : PolicySeq (S t) n) → BestExt ps (bestExt ps)
```

We come back to this in Section 4.

2.2 Lightweight vs. full theory

As noted in the beginning of this section, the theory proposed in [BJI17a] is slightly more general than the one presented here.

[5] Note that in control theory *backward induction* is often referred to as *the dynamic programming algorithm* where the term *dynamic programming* is used in the original sense of [Bel57].

In particular, in the lightweight theory, policies are just functions from states to controls:

```
Policy : (t : ℕ) → Type
Policy t = (x : X t) → Y t x
```

By contrast, in [BJI17a], policies are indexed over a number of decision steps n

```
Policy : (t, n : ℕ) → Type
Policy t Z     = Unit
Policy t (S m) = (x : X t) → Reachable x → Viable (S m) x → GoodCtrl t x m
```

and their domain for $n > 0$ is restricted to states that are *reachable* and *viable* for n steps. This allows to cope with states whose control set is empty and with transition functions that return empty M-structures of next states. (For a discussion of reachability and viability see [BJI17a, Sec. 3.7 and 3.8].)

This generality, however, comes at a cost: Compare e.g. the proof of Bellman's principle in [BB21, Appendix 5] with the corresponding proof in [BJI17a, Appendix B]. The impact of the reachability and viability constraints on other parts of the theory is even more severe.

Here, we have decided to trade some generality for better readability and ease of use, opting for a lightweight version of the original theory. Still, for the generic backward induction algorithm we need to make sure that it is possible to define policy sequences of the length required for a specific SDP. This can e.g. be done by postulating controls to be non-empty:

```
notEmptyY : (t : ℕ) → (x : X t) → Y t x
```

We also impose a non-emptiness requirement on the transition function *next* that is relevant in the context of the correctness result of [BB21] (see discussion in Section 7 of the paper).

```
nextNonEmpty : {t : ℕ} → (x : X t) → (y : Y t x) → NonEmpty (next t x y)
```

Above we have not discussed under which conditions one can implement optimal extensions of arbitrary policy sequences. This is an interesting topic but not central to the purpose of the current report. For the same reason we have not addressed the question of how to make *bi* more efficient by tabulation. We briefly discuss the specification and implementation of optimal extensions in the framework in [BB21, Appendix 7]. We refer the reader interested in tabulation of *bi* to SequentialDecisionProblems.TabBackwardsInduction of [B$^+$21].

3 Trajectories, flow and total rewards

In the previous section, we have seen the core components of the framework. In this section we introduce some additional infrastructure for working with sequential decision processes and problems.

State trajectories. Given a policy sequence (optimal or not) and an initial state for an SDP, we can compute the M-structure of possible trajectories starting at that state:

```
data StateSeq : (t, n : ℕ) → Type where
  Nil  : {t : ℕ} → StateSeq t Z
  (#)  : {t, n : ℕ} → X t → StateSeq (S t) n → StateSeq t (S n)


trjX : Monad M ⇒ {t, n : ℕ} → PolicySeq t n → X t → M (StateSeq t (S n))
trjX {t} Nil x        = pure (x # Nil)
trjX {t} (p :: ps) x = let y = p x in
                       let mx′ = next t x y in
                         map (x#) (mx′ >>= trjX ps)
```

In order to extract specific information from a state trajectory, it is helpful to have a map-like function

```
mapX : {t, n : ℕ} → {O : Type} → ({t' : ℕ} → X t' → O)
   → StateSeq t n → Vect n O
mapX obs Nil       = []
mapX obs (x # xs) = obs x :: mapX obs xs
```

Or if we explicitly need the value of t:

```
mapX' : {t, n : ℕ} → {O : Type} → ((t' : ℕ) → X t' → O)
   → StateSeq t n → Vect n O
mapX'       obs Nil       = []
mapX' {t} obs (x # xs) = obs t x :: mapX' obs xs
```

Consider a type of policy functions

```
PolicyFunction  :  Type
PolicyFunction = (t : ℕ) → (x : X t) → Y t x
```

that is, a policy function computes a control for any time step and state at that step. Then, omitting policy sequences, we can compute trajectories by

```
trjX' : Monad M ⇒ {t : ℕ} → PolicyFunction → (n : ℕ) → X t → M (StateSeq t (S n))
trjX'       _pol Z       x = pure (x # Nil)
trjX' {t} pol  (S n) x = let y    = pol t x     in
                             let mx' = next t x y in
                               map (x#) (mx' ≫= trjX' pol n)
```

We could also consider any finite trajectories as initial segment of an infinite one. Consider for this infinite sequences (streams) of states:

```
codata StateStream :(ℕ → Type) where
  XCons :{t : ℕ} → X t → StateStream (S t) → StateStream t
```

Then we can compute infinite trajectories (in principle though not in practice) by

```
partial
trjX_MStream : Monad M ⇒ PolicyFunction → (t : ℕ) → X t → M (Inf (StateStream t))
trjX_MStream pol t x = let y          = pol t x                               in
                         let mx'       = next t x y                            in
                         let mtrj      = trjX_MStream pol (S t)                in
                         let mapCons = map {f = M} (Delay ∘ XCons {t} x) in
                           mx' ≫= (mapCons ∘ mtrj)
```

However, Idris will not certify that this function is total as the definition of *trjX_MStream* does not meet the totality checker's syntactic guard condition for the recursive call.

For a deterministic transition function

```
step :(t : ℕ) → (x : X t) → Y t x → X (S t)
```

there is no such problem:

```
trjX_Stream : PolicyFunction → (t : ℕ) → X t → StateStream t
trjX_Stream pol t x = let y  = pol t x     in
                        let x' = step t x y in
                          x `XCons` trjX_Stream pol (S t) x'
```

To recover finite finite trajectories, we can use a helper function

```
take  :  (t, n : ℕ) → Inf (StateStream t) → StateSeq t n
take _t Z      _xs              = Nil
take t  (S n) (XCons x xs) = x # (take (S t) n xs)
```

such that

```
trjXLemma : Monad M ⇒ (t, n : ℕ) → (pol : PolicyFunction) → (x : X t) →
                  map (take t (S n)) (trjX_MStream pol t x) = trjX' pol n x
```

Flow. If we are not interested in the whole trajectory, but only in the final state, we can define a multi-step *flow* function from the single-step transition function *next*:

```
flow : Monad M ⇒ {t, n : ℕ} → PolicySeq t n → X t → M (X (n + t))
flow Nil x                     = pure x
flow {t} {n = S n} (p :: ps) x = rewrite plusSuccRightSucc n t in
                                   next t x (p x) >>= flow ps
```

where the *plusSuccRight n t* is the lemma $S\ (n + t) = n + S\ t$ which is required to change make the outcome type correct: The *rewrite* mechanism changes the type $M\ (X\ (S\ (n + t)))$ of *next t x (p x)* ⋙= *flow ps* (which is inferred by the Idris type checker) to the expected result type $M\ (X\ S\ (n + t))$ (which is definitionally equal[6] to $M\ (X\ (S\ n + t)))$.

More informative trajectories. Computing a "trajectory of rewards" requires more information than just a sequence of states, since rewards may depend not only on one state but also on its successor state and the control used. We therefore define more expressive trajectories

```
data StateCtrlSeq : (t, n : ℕ) → Type where
  Last : {t : ℕ} → X t → StateCtrlSeq t (S Z)
  (##) : {t, n : ℕ} → (x : X t ** Y t x) → StateCtrlSeq (S t) (S n) → StateCtrlSeq t (S (S n))

trj : Monad M ⇒ {t, n : ℕ} → PolicySeq t n → X t → M (StateCtrlSeq t (S n))
trj {t} Nil x       = pure (Last x)
trj {t} (p :: ps) x = let y = p x in
                      let mx' = next t x y in
                        map ((x ** y)##) (mx' >>= trj ps)
```

where we use *StateCtrlSeq* as type of trajectories. Essentially it is a non-empty list of (dependent) state/control pairs, with the exception of the base case which is a singleton just containing the last state reached. Now we can now compute lists of rewards

```
trjR : {t, n : ℕ} → StateCtrlSeq t n → List Val
trjR {t} (Last x)          = [zero]
trjR {t} ((x ** y) ## xys) = reward t x y (head xys) :: trjR xys
```

where *head* is a helper function that extracts the head of a state-control sequence

```
head : {t, n : ℕ} → StateCtrlSeq t (S n) → X t
head (Last x)          = x
head ((x ** y) ## xys) = x
```

Furthermore, we can compute the *total reward* for a single trajectory, i.e. its sum of rewards:[7]

```
sumR : {t, n : ℕ} → StateCtrlSeq t n → Val
sumR {t} (Last x)          = zero
sumR {t} ((x ** y) ## xys) = reward t x y (head xys) ⊕ sumR xys
```

By mapping *sumR* onto an M-structure of trajectories, we obtain an M-structure containing the individual sums of rewards of the trajectories. Now, using the measure function, we can compute the *measured total reward* for a policy sequence *ps* and an initial state x:

```
val' :  Monad M ⇒ {t, n : ℕ} → (ps : PolicySeq t n) → (x : X t) → Val
val' ps = meas ∘ map sumR ∘ trj ps
```

The measured total reward is the generic analogue of what is called the *expected total reward* in the standard case of stochastic SDPs using the expected value measure (see [Put14, ch. 4.1.2]). This function plays a crucial role in the semantic verification of the framework that has been presented in [BB21] and will be discussed in the next section.

[6] If two expressions are *definitionally equal*, the type checker can automatically transform them into the same term and thus automatically infer that they are equal, without requiring an additional equality proof like the *plusSuccRight* lemma.

[7] More concisely *sumR* may be expressed using the *foldr* operator: *sumR* = *foldr* (⊕) *zero* ∘ *trR*.

Comparison experiments. Based on the above functions for calculating trajectories and values, further infrastructure for different kinds of comparison experiments can be defined in a straightforward manner. We give here some examples.

Given a vector of policy sequences, we can compute the vector of corresponding monadic structures of trajectories:

```
trjPar  : Monad M ⇒ {n, m, t : ℕ} → Vect m (PolicySeq t n) → X t →
             Vect m (M (StateSeq t (S n)))
trjPar pss x₀ = map (flip trjX x₀) pss
```

or for a sequence of states (i.e. starting at increasing time steps) and a policy sequence:

```
trjSeq : Monad M ⇒ {t, n, m : ℕ} → ({t' : ℕ} → PolicySeq t' n) →
          StateSeq t m → Vect m (t'' : ℕ ** M (StateSeq t'' (S n)))
trjSeq {t} ps Nil        = []
trjSeq {t} ps (x # xs) = (t ** trjX ps x) :: trjSeq ps xs
```

A version of *trjPar* adding information about the initial time step to the output:

```
trjPar' : Monad M ⇒ {t, n, m : ℕ} → Vect m ((t' : ℕ) → PolicySeq t' n) →
           X t → Vect m (t : ℕ ** M (StateSeq t (S n)))
trjPar' {t} {n} []           x = []
trjPar' {t} {n} (ps :: pss) x = (t ** trjX (ps t) x) :: trjPar' pss x
```

Compute the trajectories for a vector of policy sequences and a state sequence of initial states:

```
trjParSeq : Monad M ⇒ {t, n, m, s : ℕ} → Vect s ((t' : ℕ) → PolicySeq t' n) →
             StateSeq t m → Vect m (Vect s (t : ℕ ** M (StateSeq t (S n))))
trjParSeq pss xs = mapX (trjPar' pss) xs
```

And the same for multiple state sequences of initial states:

```
trjParVSeq : Monad M ⇒ {t, n, m, s, u : ℕ} →
               Vect s ((t' : ℕ) → PolicySeq t' n) →
               Vect u (M (StateSeq t m)) →
               Vect u (M (Vect m (Vect s (t : ℕ ** M (StateSeq t (S n))))))
trjParVSeq {n} pss mxss = map (map (trjParSeq {n} pss)) mxss
```

This function may be used for a two-stage computation as indicated in Figure 2

```
twoStageAssessment : Monad M ⇒ {t, n, m, u, v : ℕ} →
                        (pss : Vect u (PolicySeq t n)) →
                        (pss' : Vect v ((t' : ℕ) → PolicySeq t' m)) → X t
                         → Vect u (M (Vect (S n) (Vect v (t : ℕ ** M (StateSeq t (S m))))))
twoStageAssessment {t} pss pss' =
  let stage1 = trjPar pss in
  let stage2 = trjParVSeq pss' in
    stage2 ∘ stage1
```

This kind of assessment will be relevant for the notion of *lost option commitment* explored in the context of TiPES Deliverable D6.3 [MMCBB22a, MMCBB22b]. If we are not interested in full (monadic) trajectories but only in the final states or the aggregated values resulting from different policy sequences, we can use the following functions:

```
flowPar : Monad M ⇒ {t, n, m : ℕ} → Vect m ((t' : ℕ) → PolicySeq t' n) →
           X t → Vect m (M (X (n + t)))
flowPar {t} {n} []           x = []
flowPar {t} {n} (ps :: pss) x = flow (ps t) x :: flowPar pss x
```

```
flowParSeq : Monad M ⇒ {t, n, m, s : ℕ} → StateSeq t n →
               Vect s ((t' : ℕ) → PolicySeq t' m) →
```

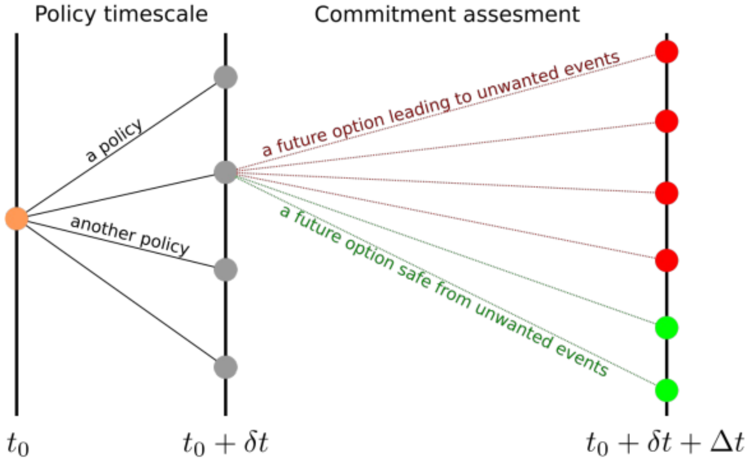

Policy timescale
Commitment assesment
a policy
another policy
a future option leading to unwanted events
a future option safe from unwanted events
t_0
$t_0 + \delta t$
$t_0 + \delta t + \Delta t$

Figure 2: Two-stage parallel computation along multiple policy sequences illustrating the computation of *lost option commitment* explored in the context of TiPES Deliverable D6.3 [MMCBB22a, MMCBB22b].

```
                   Vect n (t'' ** Vect s (M (X (m + t''))))
flowParSeq xs pss = mapX' (λt, x ⇒ (t ** (flowPar pss x))) xs
```

and

```
valPar : Monad M ⇒ {t, n, m : ℕ} → Vect m ((t' : ℕ) → PolicySeq t' n) →
         X t → Vect m Val
valPar {t} {n} []          x = []
valPar {t} {n} (ps :: pss) x = val (ps t) x :: valPar pss x

valParSeq : Monad M ⇒ {t, n, m, s : ℕ} →
            Vect s ((t' : ℕ) → PolicySeq t' m) →
            StateSeq t n → Vect n (Vect s Val)
valParSeq pss xs = mapX (λx ⇒ valPar pss x) xs

twoStageValuation : Monad M ⇒ {t, n, m, u, v : ℕ} →
                    (pss : Vect u (PolicySeq t n)) →
                    (pss' : Vect v ((t' : ℕ) → PolicySeq t' m)) →
                    X t → Vect u (M (Vect (S n) (Vect v Val)))
twoStageValuation {u} pss pss' =
  let stage1 = trjPar pss                  in
  let stage2 = map (map (valParSeq pss')) in
    stage2 ∘ stage1
```

Based on the function introduced in this section we can describe typical experiments like sensitivity or commitment computations [BBCMM22c, Section 3+4].

Numerical simulation. As can be seen from the type of the *next* function, the basic formalism does assume that a time-discrete one step transition function is given. Although numerical simulation is not the main objective of the framework, we sketch how to arrive at a sequential decision process when starting from a ordinary differential equation (ODE) description of a parameterised and forced dynamical system.

Consider an ODE

$$\frac{dx}{dt}(t) = f(p, t, x(t))$$

where $t : \mathcal{T}$, $x : \mathcal{T} \to \mathbb{R}^n$ for some $n : \mathbb{N}$, and $p : P$ where P is a type of parameters. The function f defining the right-hand side of the ODE has the type $P \times \mathcal{T} \times \mathbb{R}^n \to \mathbb{R}^n$. Typically one has on the one hand constant parameters, on the other hand time-dependent forcings. In implementations, $\mathbb{R}$ is typically represented by floating point numbers. He we follow this common approach and do not go into formalisation of real number arithmetic.

Using slightly more general types than above, let f be encoded as a function of type

```
Time :  Type
Time = Double

Rn : Type
P  : Type

f :    P → Time → Rn → Rn
```

Given this representation of the RHS of the ODE, the simplest way to obtain a one step function is the forward Euler method using a uniform time step *delta_t* : *TimeDiff*

```
TimeDiff  : Type
TimeDiff = Double

delta_t : TimeDiff
```

an addition

```
add : Rn → Rn → Rn
```

and a scalar multiplication

```
smult : TimeDiff → Rn → Rn


eulerForward : {X, X', Param : Type} →
                      (add : X → X' → X) → (smult : TimeDiff → X' → X') →
                      (rhs : Param → Time → X → X') → TimeDiff → Param → Time → X → X
eulerForward add smult rhs dt p t x = x `add` (dt `smult` (rhs p t x))
```

Given a function *natToTime* that associates decision steps with points in time

```
natToTime : ℕ → Time
```

one might thus define a deterministic sequential decision process with

```
Framework.M T = T
X _           = Rn
Y _ _         = P


next t x y = eulerForward add smult f delta_t y (natToTime t) x
```

Now, for a real application, one would of course want to do a number of improvements of the basic idea described above. Though not in the scope of the current report, especially more accurate integration methods would be important, and could be supported by a convenient DSL to derive decision processes from ODEs.

However, a shortcoming of the above that is easy to fix is the following: For a decision problem on top of such a process, one would very likely not want to associate one decision step with one integration step.

This can be easily solved by using a generic iteration function similar to the flow function defined above:

```
iter : {A : Type} → (Time → A → A) → TimeDiff → ℕ → Time → A → A
iter f dt Z      t a = a
iter f dt (S n) t a = iter f dt n (t + dt) (f t a)
```

Now, if one step of next should amount to e.g. 1000 integration steps, one can define *next* by

```
next t x y = let aux = eulerForward add smult f delta_t y in
                 iter aux delta_t 1000 (natToTime t) x
```

Note that the value of the parameter is kept constant during the intermediate iterations. This is coherent with the idea of a control parameter that can only be influenced at the decision steps. However, if the control parameters are meant to represent a continuous function, this might not be the intended behaviour. In this case it seems reasonable to use some form of interpolation to recover intermediate values and make the resulting function part of the parameters of the system.

4 Correctness of the framework

An important aspect of using a dependently typed programming language is that it allows to specify and implement programs and prove their correctness with respect to the specification all in the same language.

In [BJI17a], it was formally proved that the backward induction of the full theory computes policy sequences that are optimal with respect to the value function *val* of section 2.1 which is a monadic generalisation of the Bellman equation [Bel57]. However, in the literature on stochastic SDPs this formulation of the value function is itself part of the backward induction algorithm and needs to be verified against an *objective function* or *optimisation criterion*, called the *expected total reward* in [Put14, Ch. 4.2]. For stochastic SDPs semi-formal proofs can be found in textbooks – but monadic SDPs are substantially more general than the stochastic SDPs for which these results are established. This observation raises a number of questions:

- What exactly should "correctness" mean for a solution of monadic SDPs?
- Does monadic backward induction provide correct solutions in this sense for monadic SDPs in their full generality?
- And if not, is there a class of monadic SDPs for which monadic backward induction does provide provably correct solutions?

We have addressed these questions in [BB21] and made the following contributions to answering them:

- We put forward a formal specification that monadic backward induction should meet in order to be considered "correct" as solution method for monadic SDPs. This specification uses an optimisation criterion that is a generic version of the *expected total reward* of standard control theory textbooks. In analogy, we call this criterion *measured total reward* (computed by the function *val′* defined in Section 3).
- We consider the value function underlying monadic backward induction as "correct" if it computes the *measured total reward.*
- If the value function *val* is correct in this sense, then monadic backward induction can be proven to be correct by extending the result of [BJI17a]. However, we showed in [BB21] that *val* does not compute the *measured total reward* for arbitrary monadic SDPs that only fulfil the axioms of the [BJI17a] theory.
- We therefore formulated compatibility conditions that identify a class of monadic SDPs for which *val* and thus monadic backward induction can be shown to be correct. The conditions (*measPureSpec*, *measJoinSpec* and *measPlusSpec* of Sectionsubsection:framework:specsol) are fairly simple and allow for a neat description in category-theoretical terms using the notion of Eilenberg-Moore-algebra.
- We gave a formalised proof that monadic backward induction fulfils the correctness criterion if the conditions hold. This correctness result can be seen as a generic version of correctness results for standard backward induction like [Ber95, Prop. 1.3.1] and [Put14, Th. 4.5.1.c].

Below we will outline the two parts of the correctness proof for the lightweight theory of Section 2.1. The full proofs, discussion of the compatibility conditions and reasons why they are required can be found in [BB21]. It is worth stressing that our conditions can be useful for anyone interested in applying monadic backward induction in non-standard situations – completely independent of the BJI-framework.

4.1 Specifying correctness

By analogy to the case of stochastic SDPs treated in textbooks like [Put14], we define backward induction for monadic SDPs to be *correct* if it computes a policy sequence that results in the optimal measured total reward for any initial state. This is expressed by the following specification:

$$\mathit{biOptMeasTotalReward} : \mathit{Monad}\ M \Rightarrow (t, n : \mathbb{N}) \to \mathit{OptimalPolicySeq}\ \mathit{val}'\ (\mathit{bi}\ t\ n)$$

where *OptimalPolicySeq* is an generic optimality predicate that takes as parameters an objective function and a policy sequence. The predicate holds if the policy sequence is optimal with respect to the objective function. Above, the objective function is the measured total reward val'.

$$\begin{aligned}&\mathit{OptimalPolicySeq} : \{t, n : \mathbb{N}\} \to (\mathit{PolicySeq}\ t\ n \to X\ t \to \mathit{Val}) \to \mathit{PolicySeq}\ t\ n \to \mathit{Type}\\ &\mathit{OptimalPolicySeq}\ \{t\}\ \{n\}\ f\ \mathit{ps} = (\mathit{ps}' : \mathit{PolicySeq}\ t\ n) \to (x : X\ t) \to f\ \mathit{ps}'\ x \sqsubseteq f\ \mathit{ps}\ x\end{aligned}$$

In [BJI17a], Botta *et al.* have shown (for the full theory) that if M is a monad, $\sqsubseteq$ a total preorder and $\oplus$ and *meas* fulfil the two monotonicity conditions *measMon* and *plusMon*, then $\mathit{bi}\ t\ n$ yields an optimal policy sequence with respect to the value function *val* in the sense that $\mathit{val}\ \mathit{ps}'\ x \sqsubseteq \mathit{val}\ (\mathit{bi}\ t\ n)\ x$ for any policy sequence ps' and initial state x, for any $t, n : \mathbb{N}$. Or, expressed using the generic optimality predicate, that the type

$$\mathit{OptimalPolicySeq}\ \{t\}\ \{n\}\ \mathit{val}\ (\mathit{bi}\ t\ n)$$

is inhabited. As seen in Sec. 2.1, the function *val* measures and adds rewards incrementally. But does it always compute the measured total reward like val'? Modulo differences in the presentation [Put14, Theorem 4.2.1] suggests that for standard stochastic SDPs, *val* and val' are extensionally equal, which in turn allows the use of backward induction for solving these SDPs. Generalising, we therefore consider *val* as correct if it fulfils the specification

$$\mathit{valMeasTotalReward} : \{t, n : \mathbb{N}\} \to (\mathit{ps} : \mathit{PolicySeq}\ t\ n) \to (x : X\ t) \to \mathit{val}\ \mathit{ps}\ x = \mathit{val}'\ \mathit{ps}\ x$$

If this equality holds for the general monadic SDPs of the framework, we can prove the correctness of *bi* as immediate corollary of *valMeasTotalReward*. We therefore proceed as follows:

(1.) Prove $\mathit{OptimalPolicySeq}\ \mathit{val}\ (\mathit{bi}\ t\ n)$: *bi* computes optimal policy sequences wrt *val*

(2.) Prove *valMeasTotalReward*: *val* is extensionally equal to val'

(3.) Deduce *biOptMeasTotalReward*: *bi* computes optimal policy sequences wrt val'

4.2 Optimality with respect to *val*

The generic implementation of backward induction of the [BJI17a] theory uses a generalisation of *Bellman's principle of optimality*. In control theory textbooks, this principle is often referred to as *Bellman's equation*. It can be suitably formulated in terms of the notion of *optimal extension*. Recall from Section 2.1 that we say that a policy $p : \mathit{Policy}\ t$ is an optimal extension of a policy sequence $\mathit{ps} : \mathit{Policy}\ (S\ t)\ n$ if it is the case that the value of $p :: \mathit{ps}$ is at least as good as the value of $p' :: \mathit{ps}$ for any policy p' and for any state $x : X\ t$:

$$\begin{aligned}&\mathit{BestExt} : \mathit{Functor}\ M \Rightarrow \{t, n : \mathbb{N}\} \to \mathit{PolicySeq}\ (S\ t)\ n \to \mathit{Policy}\ t \to \mathit{Type}\\ &\mathit{BestExt}\ \{t\}\ \mathit{ps}\ p = (p' : \mathit{Policy}\ t) \to (x : X\ t) \to \mathit{val}\ (p' :: \mathit{ps})\ x \sqsubseteq \mathit{val}\ (p :: \mathit{ps})\ x\end{aligned}$$

With the notion of optimal extension in place, Bellman's principle can then be formulated as

```
Bellman : Functor M ⇒ {t, n : ℕ} →
          (ps : PolicySeq (S t) n) → OptimalPolicySeq val ps →
          (p  : Policy t) → BestExt ps p →
          OptimalPolicySeq val (p :: ps)
```

In words: *extending an optimal policy sequence with an optimal extension (of that policy sequence) yields an optimal policy sequence.* Another way of expressing the same principle is to say that prefixing with optimal extensions preserves optimality.

Proving Bellman's optimality principle is almost straightforward and crucially relies on $\sqsubseteq$ being reflexive and transitive (remember that $\sqsubseteq$ is a total preorder). The proof obligation is to show that

```
val (p′ :: ps′) x ⊑ val (p :: ps) x
```

for arbitrary p', ps' and x of suitable types. This is achieved by transitivity of $\sqsubseteq$ on two sub proofs:

```
val (p′ :: ps′) x ⊑ val (p′ :: ps) x
```

and

```
val (p′ :: ps) x ⊑ val (p :: ps) x
```

The first inequality follows from the optimality of *ps* (the second argument of Bellman), reflexivity of $\sqsubseteq$ and from the two *monotonicity* properties *plusMon* and *measMon* from Section 2.1:

```
plusMon  : {v1, v2, v3, v4 : Val} →
           v1 ⊑ v2 → v3 ⊑ v4 → (v1 ⊕ v3) ⊑ (v2 ⊕ v4)
measMon : Functor M ⇒ {A : Type} →
           (f, g : A → Val) → ((a : A) → f a ⊑ g a) →
           (ma : M A) → meas (map f ma) ⊑ meas (map g ma)
```

The condition *measMon* is a special case of the measure monotonicity requirement originally formulated by C. Ionescu in [Ion09] in the framework of a theory of vulnerability and monadic dynamical systems. It is a natural property that, among others, the expected value measure and the worst (best) case measure do fulfil.

The second inequality directly follows from the last argument of Bellman, a proof that *BestExt ps p*. We provide a proof of *Bellman* in [BB21, Appendix 5]. As one would expect, the proof crucially depends on the recursive definition of *val* discussed above. With *Bellman* in place, we can straightforwardly prove that the generic monadic implementation of backward induction *bi* computes policy sequences that are optimal with respect to *val*. The proof is by induction on the length of the policy sequence.

For the base case, note that the empty policy sequence is optimal because any empty policy sequence will result in value *zero* and $\geqslant$ is reflexive:

```
biLemmaBase : Functor M ⇒ OptimalPolicySeq val Nil
biLemmaBase Nil x = reflexive lteTP zero
```

In the step case, we apply Bellman's principle. We thus get the following proof of optimality with respect to *val*:

```
biLemma : Functor M ⇒ (t : ℕ) → (n : ℕ) → OptimalPolicySeq val (bi t n)
biLemma t Z      = biLemmaBase   -- base case
biLemma t (S n) =                -- step case
  let ps  = bi (S t) n         in  -- ps computed by backward induction
  let ops = biLemma (S t) n in  -- induction hypothesis
  let p   = bestExt ps         in  -- p computed by best extension function
  let oep = bestExtSpec ps   in  -- specification of best extension
    Bellman ps ops p oep         -- application of Bellman's principle
```

4.3 Extensional equality of *val* and *val′*

Now we show that the monadic value function *val* based on Bellman's equation computes the measured total reward by showing that the functions *val* and *val′* are extensionally equal

$$\mathit{valMeasTotalReward} : \{\, t, n : \mathbb{N} \} \to (\mathit{ps} : \mathit{PolicySeq}\ t\ n) \to (x : X\ t) \to \mathit{val}'\ \mathit{ps}\ x = \mathit{val}\ \mathit{ps}\ x$$

The proof of *valMeasTotalReward* is slightly more involved than the proof of *biLemma*. Therefore we will present the main part of the proof in semi-formal equational style here, while the full implementation can be found in [BB21, Appendix 2].

Inspecting the definitions of *val* and *val′*, we see that they exhibit different computational patterns: While *val′* first computes all possible trajectories for the given policy sequence and initial state, then computes their individual sum of rewards and finally applies the measure once, *val* computes its final result by adding the current reward to an intermediate outcome and applying the measure locally at each decision step. This suggests that a transformation from *val′* to *val* will essentially have to push the application of the measure into the recursive computation of the sum of rewards. The proof is carried out by induction on the structure of policy sequences. It hinges on the three compatibility conditions between the monad M, the measure *meas* and the operation $\oplus$ stated in Section 2.1, *measPure*, *measJoin* and *measPlus*. These properties ensure that the objective function *val′* can be transformed into the more efficient *val* without loss of information.[8]

Lemmas. Based on the general functor and monad properties of M and the compatibility conditions, we can prove the following technical lemmas:

$$\begin{aligned}
&\mathit{measAlgLemma} &&: \{A, B : \mathit{Type}\} \to (f : B \to \mathit{Val}) \to (g : A \to M\ B) \to \\
& &&\quad (\mathit{meas} \circ \mathit{map}\ (\mathit{meas} \circ \mathit{map}\ f \circ g)) \doteq (\mathit{meas} \circ \mathit{map}\ f \circ \mathit{join} \circ \mathit{map}\ g) \\
&\mathit{headTrjLemma} &&: \{\, t, n : \mathbb{N}\} \to (\mathit{ps} : \mathit{PolicySeq}\ t\ n) \to (r : X\ t \to \mathit{Val}) \to \\
& &&\quad (s : \mathit{StateCtrlSeq}\ t\ (S\ n) \to \mathit{Val}) \to (x : X\ t) \to \\
& &&\quad (\mathit{map}\ (r \circ \mathit{head} \bigoplus s) \circ \mathit{trj}\ \mathit{ps})\ x = \\
& &&\quad (\mathit{map}\ (\mathit{const}\ (r\ x) \bigoplus s) \circ \mathit{trj}\ \mathit{ps})\ x \\
&\mathit{measSumLemma} &&: \{\, t, n : \mathbb{N}\} \to (\mathit{ps} : \mathit{PolicySeq}\ t\ n) \to \\
& &&\quad (r : X\ t \to \mathit{Val}) \to \\
& &&\quad (s : \mathit{StateCtrlSeq}\ t\ (S\ n) \to \mathit{Val}) \to \\
& &&\quad (\mathit{meas} \circ \mathit{map}\ (r \circ \mathit{head} \bigoplus s) \circ \mathit{trj}\ \mathit{ps}) \doteq \\
& &&\quad (r \bigoplus \mathit{meas} \circ \mathit{map}\ s \circ \mathit{trj}\ \mathit{ps})
\end{aligned}$$

The first lemma allows us to lift and eliminate an application of the monad's *join* operation. [9] The second lemma says that mapping the function *head* onto an M-structure of trajectories computed with *trj* results in an M-structure filled with the initial states of these trajectories. The third lemma allows us to both commute the measure into the right summand of an $\bigoplus$-sum and to perform the head/trajectory simplification. It lies at the core of the relationship between *val* and *val′*.

Main proof. With these lemmas in place, we can prove that *val* is extensionally equal to *val′*.

Let $t, n : \mathbb{N}$, $\mathit{ps} : \mathit{PolicySeq}\ t\ n$. We prove *valMeasTotalReward* by induction on *ps*.

Base case. We need to show that for all $x : X\ t$, $\mathit{val}'\ \mathit{Nil}\ x = \mathit{val}\ \mathit{Nil}\ x$. The right hand side of this equation reduces to *zero* by definition. The left hand side can be simplified to *meas* (*pure zero*) since *pure* is a natural transformation. At this point, our first condition, *measPureSpec*, comes into play: Using that *meas* is inverse to *pure* on the left, we can conclude that the equality holds.

In equational reasoning style: For all $x : X\ t$,

[8] Essentially, the proof consists of an algorithm that performs this transformation.

[9] This lemma is generic in the sense that it holds for arbitrary Eilenberg-Moore algebras of a monad. Here we prove it for the framework's measure *meas*, but note that in [BB21, Appendix 4.1] we prove a generic version that is then appropriately instantiated.

$$
\begin{array}{ll}
\multicolumn{2}{l}{\mathit{valMeasTotalReward}\ \mathit{Nil}\ x =} \\
\quad (\mathit{val}'\ \mathit{Nil}\ x) & =\{\text{ by definition of } \mathit{val}'\ \}= \\
\quad (\mathit{meas}\ (\mathit{map}\ \mathit{sumR}\ (\mathit{trj}\ \mathit{Nil}\ x))) & =\{\text{ by definition of } \mathit{trj}\ \}= \\
\quad (\mathit{meas}\ (\mathit{map}\ \mathit{sumR}\ (\mathit{pure}\ (\mathit{Last}\ x)))) & =\{\ \mathit{pure} \text{ is a natural transformation }\}= \\
\quad (\mathit{meas}\ (\mathit{pure}\ (\mathit{sumR}\ (\mathit{Last}\ x)))) & =\{\text{ by definition of } \mathit{sumR}\ \}= \\
\quad (\mathit{meas}\ (\mathit{pure}\ \mathit{zero})) & =\{\text{ by } \mathit{measPureSpec}\ \}= \\
\quad (\mathit{zero}) & =\{\text{ by definition of } \mathit{val}\ \}= \\
\quad (\mathit{val}\ \mathit{Nil}\ x) &
\end{array}
$$

Step case. The induction hypothesis (IH) is: for all $x : X\ t$, $\mathit{val}'\ \mathit{ps}\ x = \mathit{val}\ \mathit{ps}\ x$. We have to show that IH implies that for all $p : \mathit{Policy}\ t$ and $x : X\ t$, the equality $\mathit{val}'\ (p :: \mathit{ps})\ x = \mathit{val}\ (p :: \mathit{ps})\ x$ holds. For brevity (and to economise on brackets), let in the following $y = p\ x$, $\mathit{mx}' = \mathit{next}\ t\ x\ y$, $r = \mathit{reward}\ t\ x\ y$, $\mathit{trjps} = \mathit{trj}\ \mathit{ps}$, and $\mathit{consxy} = ((x \ast\ast y) \#\#)$.

As in the base case, all that has to be done on the *val*-side of the equation only depends on definitional equality. However, it is more involved to bring the *val′*-side into a form in which the induction hypothesis can be applied. This is where we leverage on the lemmas proved above.

By definition and because *map* preserves composition, we know that $\mathit{val}'\ (p :: \mathit{ps})\ x$ is equal to $(\mathit{meas} \circ \mathit{map}\ ((r \circ \mathit{head}) \bigoplus \mathit{sumR}))\ (\mathit{mx}' \gg\!= \mathit{trjps})$. We use the relation between the monad's *bind* and *join* to eliminate the *bind*-operator from the term. Now we can apply the first lemma from above, *measAlgLemma*, to lift and eliminate the *join* operation.

To commute the measure under the $\bigoplus$ and get rid of the application of *head*, we use our third lemma, *measSumLemma*. At this point we can apply the induction hypothesis and the resulting term is equal to $\mathit{val}\ \mathit{ps}\ x$ by definition.

The more detailed equational reasoning proof:

$$
\begin{array}{ll}
\multicolumn{2}{l}{\mathit{valMeasTotalReward}\ (p :: \mathit{ps})\ x =} \\
\quad (\mathit{val}'\ (p :: \mathit{ps})\ x) & =\{\text{ by definition of } \mathit{val}'\ \}= \\
\quad (\mathit{meas}\ (\mathit{map}\ \mathit{sumR}\ (\mathit{trj}\ (p :: \mathit{ps})\ x))) & =\{\text{ by definition of } \mathit{trj}\ \}= \\
\quad (\mathit{meas}\ (\mathit{map}\ \mathit{sumR}\ (\mathit{map}\ \mathit{consxy}\ (\mathit{mx}' \gg\!= \mathit{trjps})))) & =\{\ \mathit{map} \text{ preserves composition }\}= \\
\quad (\mathit{meas}\ (\mathit{map}\ (\mathit{sumR} \circ \mathit{consxy})\ (\mathit{mx}' \gg\!= \mathit{trjps}))) & =\{\text{ by definition of } \mathit{sumR}\ \}= \\
\quad (\mathit{meas}\ (\mathit{map}\ ((r \circ \mathit{head}) \bigoplus \mathit{sumR})\ (\mathit{mx}' \gg\!= \mathit{trjps}))) & =\{\text{ relation } \mathit{bind}/\mathit{join}\ \}= \\
\quad (\mathit{meas}\ (\mathit{map}\ ((r \circ \mathit{head}) \bigoplus \mathit{sumR})\ (\mathit{join}\ (\mathit{map}\ \mathit{trjps}\ \mathit{mx}')))) & =\{\text{ by } \mathit{measAlgLemma}\ \}= \\
\quad (\mathit{meas}\ (\mathit{map}\ (\mathit{meas} \circ \mathit{map}\ (r \circ \mathit{head} \bigoplus \mathit{sumR}) \circ \mathit{trjps})\ \mathit{mx}')) & =\{\text{ by } \mathit{measSumLemma}\ \}= \\
\quad (\mathit{meas}\ (\mathit{map}\ (r \bigoplus \mathit{meas} \circ \mathit{map}\ \mathit{sumR} \circ \mathit{trjps})\ \mathit{mx}')) & =\{\text{ by definition of } \mathit{val}'\ \}= \\
\quad (\mathit{meas}\ (\mathit{map}\ (r \bigoplus \mathit{val}'\ \mathit{ps})\ \mathit{mx}')) & =\{\text{ by induction hypothesis }\}= \\
\quad (\mathit{meas}\ (\mathit{map}\ (r \bigoplus \mathit{val}\ \mathit{ps})\ \mathit{mx}')) & =\{\text{ by definition of } \mathit{val}\ \}= \\
\quad (\mathit{val}\ (p :: \mathit{ps})\ x) &
\end{array}
$$

□

Technical remarks. The above proof of *valMeasTotalReward* omits some technical details that may be uninteresting for a pen and paper proof, but turn out to be crucial in the setting of an intensional type theory – like Idris – where function extensionality does not hold in general. In particular, we have to postulate that the functorial *map* preserves extensional equality (see [BB21, Appendix 1.2] and [BBJR21] for details) for Idris to accept the proof. In fact, most of the reasoning proceeds by replacing functions that are mapped onto monadic values by other functions that are only extensionally equal. Using that *map* preserves extensional equality allows to carry out such proofs generically without knowledge of the concrete structure of the functor.

4.4 Concluding correctness

Using the results from above, we can now prove the correctness of monadic backward induction, namely that the policy sequences computed by *bi* are optimal with respect to the measured total reward computed by *val′*:

```
biOptMeasTotalReward : (t, n : ℕ) → OptimalPolicySeq val′ (bi t n)
biOptMeasTotalReward t n ps′ x =
  let vvEqL = sym (valMeasTotalReward ps′ x)        in
  let vvEqR = sym (valMeasTotalReward (bi t n) x) in
  let biOpt  = biOptVal t n ps′ x                     in
  replace vvEqR (replace vvEqL biOpt)
```

The statement *biOptMeasTotalReward* can be seen as a generic version of textbook correctness statements for backward induction as solution method for stochastic SDPs like [Ber95, prop.1.3.1] or [Put14, Theorem 4.5.1.c]. By proving *valMeasTotalReward* we have therefore extended the verification of [BJI17a] and obtained a stronger correctness result for monadic backward induction.

5 Example: A GHG emission SDP

In this section we revisit the stochastic decision process underlying the SDP discussed in [BBC$^+$21]. We first describe the problem informally. Then we give a formal specification of the underlying decision process and discuss how to derive a modular definition of the transition function guided by a Bayesian network. We will come back to this example in Section 7 and extend it to a full sequential decision problem.

5.1 Informal description of the problem

Consider a GHG emissions process in which now and for a few more decades, humanity (taken here as a global decision maker) faces two options:

1. Start a “green” transition by reducing GHG emissions according to a “safe” corridor, for example, the one depicted at page 15, Figure SPM.3a of the IPCC Summary for Policymakers [Int18]

2. Delay such transition.

In other words, assume that, over the time period between two subsequent decisions (say, for concreteness, a decade), either a transition to a nearly carbonised global socio-economic system is started or nothing happens. Further, assume that, once a transition has been started, it cannot be halted or reversed by later decisions or events. We consider this oversimplified situation only for the sake of clarity, although it might well be that green transitions are in fact fast and irreversible [ODC$^+$20].

Selecting to start a green transition in a specific physical, social and economic condition yields a different “new” condition at the next decision step. Let’s call one such condition a *micro-state*.

The idea is that micro-states are detailed descriptions of physical, social and economic observables. For example, a micro-state could encode values of GHG concentrations in the atmosphere, carbon mass in the ocean upper layer, global temperature deviations, frequency of extreme events, values of economic growth indicators, measures of inequality, etc. Even if we knew the “current” micro-state perfectly, the set of possible micro-states at the next decision step (say, one decade later) would still be extremely large, reflecting both the epistemic uncertainties (imperfect knowledge) about the (physical, social and economical) processes that unfold in the time between now and the next decision step and the aleatoric uncertainty [She19] of those processes.

Descriptions of decision processes explicitly based on micro-states would be both computationally intractable and, as discussed in detail in section 6, methodologically questionable. As in the the car accident example quoted at the opening of this section, we avoid these shortcomings by considering only a small number of sets (clusters, partitions) of micro-states. These *macro-states* (in the following, just states) consist of micro-states in which:

- A green transition has been *started* or *delayed* (S-states, D-states).
- The economic wealth is *high* or *low* (H-states, L-states).
- The world is *committed* or *uncommitted* to severe CC impacts (C-states, U-states).

In other words, we only distinguish between 8 possible states: DHU, DHC, DLU, DLC, SHU, SHC, SLU and SLC where DHU represent micro-states in which a green transition has been *delayed*, economic wealth is *high* and the world is *uncommitted* to future severe CC impacts. Similarly for DHC, DLU, etc.

Clearly, this is a very crude simplification. But it is useful to study the impact of uncertainty on relevant climate decisions and sufficient to illustrate our approach towards measuring how much decisions matter. Also, notice that binary partitioning of micro-states is at the core of the original notion of planetary boundaries [RSN+09], of the topological classification proposed in [HKDM16a] and of the social dilemmas discussed in [BDLK18].

The decision process starts in DHU. In this state, a decision to start a green transition can lead to any of the $DHU \ldots SLC$ states, albeit *with different probabilities*: the idea is that the probability of reaching states in which the green transition has been started (S-states) is *higher* than the probability of reaching D-states, in which the green transition has been delayed. Symmetrically, we assume that the decision to delay the start of a green transition in DHU is more likely to yield D-states than S-states.

In other words, we assume our (global, idealised) decision maker to be *effective*, but only to a certain degree. This accounts for the fact that, in practice, decisions are not always implemented, be this because global coordination is necessarily imperfect, because global players tend to be in competition and legislation tends to have large inertia or perhaps because some other global challenge (a pandemic or an economic downturn) has taken centre stage. As demonstrated in [BJI18], limited effectiveness has a significant impact on optimal GHG emissions policies. Thus, it would be inappropriate to assume that decisions are always implemented with certainty.

Another essential trait of our stylised process is that decisions to start a green transition, if implemented, are more likely to yield states with a low level of economic wealth (L-states) than states with high economic wealth. This assumption reflects the fact that starting a green transition requires more investments and costs than just moving to states in which most of the work towards a globally decarbonised society remains to be done.

Finally, we assume that the probability of entering states in which the world is committed to severe CC impacts is higher in states in which a green transition has not already been started as compared to states in which a green transition has been started. Also, as one would expect, delaying transitions to decarbonised economies *increases* the likelihood of entering states in which the world is committed to severe CC impacts.

Next, we give a complete formal specification of our stylised decision process.

5.2 Formal specification of the decision process

Monad. As a first step, we have to define the uncertainty monad M. As discussed in the introduction, our stylised GHG emission process is a *stochastic* process and thus M represents stochastic uncertainty:

$$\mathit{Framework.M} = \mathit{SimpleProb}$$

Here, *SimpleProb* is a finite probability monad: for an arbitrary type A, a value of type $\mathit{SimpleProb}\ A$ just consists of a list of elements of type $(A, \mathit{Double}_{\geqslant 0})$ together with a proof that the sum of the seconds of such list is positive.

States and controls. Second, we have to specify the states and the controls of the process. The states informally introduced above consisted of three components: one to indicate whether a green transition had been started or not (S or D), one to represent the economy (L or H) and one to indicate whether the state is considered committed or not (U or C). Accordingly, we just define types for each of these components:

```
data SD = S | D
data LH = L | H
data UC = U | C
```

Then each informal state can be represented as a triple, e.g. *DHU* as (D, H, U). Thus, we define the type of states as[10]

```
State : Type
State = (SD, LH, UC)
X _t = State
```

Third, we have to specify the controls of the stylised GHG emission process. Above, we said that in states in which a green transition has not already been started (that is, in D-states), the decision maker has the option of either starting or further delaying a green transition.

```
data StartDelay = Start | Delay
Y _t (D, _x2, _x3) = StartDelay
```

However, if a a green transition has already been started, the decision maker has no alternatives. We formalise this idea by defining the set of controls in S-states to be a singleton:

```
Y _t (S, _x2, _x3) = Unit
```

It will be useful to have two functions that test whether a state is committed to impacts from climate change and whether the economic wealth has taken a downturn:

```
isCommitted : (t : ℕ) → X t → 𝔹
isCommitted _t (_x1, _x2, U) = False
isCommitted _t (_x1, _x2, C) = True

isDisrupted : (t : ℕ) → X t → 𝔹
isDisrupted _t (_x1, H, _x3) = False
isDisrupted _t (_x1, L, _x3)  = True
```

That is, *isCommitted* (*isDisrupted*) returns *True* in *C*-states (*L*-states) and *False* in *U*-states (*H*-states).

The *next* function. Finally, we have to specify the transition function of the sequential decision process. As discussed in the introduction, this is defined in terms of transition probabilities (all these probabilities are of type $Double_{\geqslant 0}$).

- *The probabilities of starting a green transition:*

  Let's first specify the probability that a green transition is started, conditional to the decision taken by the decision maker. Let

  $$p_{S|Start} : Double_{\geqslant 0}$$

  denote the probability that a green transition is started (during the time interval between the current and the next decision step) given that the decision maker has decided to start it. For a perfectly effective decision maker, $p_{S|Start}$ would be one.

[10] If we do not use an argument, we indicate this by prefixing the variable name by an underscore like the variable _t in the definitions of X or Y.

Let's assume a 10% chance that a decision to start a green transition fails to be implemented, perhaps because of inertia of legislation, as discussed above:

$$p_{S|Start} = \ 0.9$$

Consistently, the probability that a green transition is delayed even if the decision maker has chosen to start it is

$$\begin{aligned} &p_{D|Start} \ : \ \mathit{Double}_{\geqslant 0} \\ &p_{D|Start} = \mathit{one} - p_{S|Start} \end{aligned}$$

Similarly, we denote with $p_{D|Delay}$ and $p_{S|Delay}$ the probabilities that a green transition is delayed (started) given that the decision maker has decided to delay it. As a first step, we take $p_{S|Delay}$ to be equal to $p_{D|Start}$

$$\begin{aligned} &p_{D|Delay} \ : \ \mathit{Double}_{\geqslant 0} \\ &p_{D|Delay} = \ 0.9 \\ &p_{S|Delay} \ : \ \mathit{Double}_{\geqslant 0} \\ &p_{S|Delay} = \mathit{one} - p_{D|Delay} \end{aligned}$$

albeit we might want to .... We want to make sure that the values of $p_{S|Start}$, $p_{D|Start}$, $p_{D|Delay}$ and $p_{S|Delay}$ are consistent with the assumption (remember the informal description of our stylised GHG emission process from the introduction) that our decision maker is, up to a certain degree, effective and require them to fulfil the inequalities

$$\begin{aligned} &\mathit{pSpec1} : p_{D|Start} \leqslant p_{S|Start} \\ &\mathit{pSpec2} : p_{S|Delay} \leqslant p_{D|Delay} \end{aligned}$$

- *The probabilities of economic downturns:*

In the informal description of the decision process, we said that an essential trait of the decision process is that

> ...decisions to start a green transition, if implemented, are more likely to yield states with a low level of economic wealth (L-states) than states with high economic wealth. This assumption reflects the fact that starting a green transition requires more investments and costs than just moving to states in which most of the work towards a globally decarbonised society remains to be done.

We need to formulate this idea in terms of transition probabilities. Let $p_{L|S,DH}$ denote the probability of transitions to states with a low level of economic wealth (L) given that a green transition has been started (S) from delayed states (D) with a high level of economic wealth (H). Similar interpretations hold for $p_{L|S,DL}$, $p_{L|S,SH}$, $p_{L|S,SL}$ and their counterparts for the cases in which a green transition has been delayed, $p_{L|D,DH}$ and $p_{L|D,DL}$. Remember that in our decision process

> ...once a transition has been started, it cannot be halted or reversed by later decisions or events.

In terms of transition probabilities, this means that we do not need to specify $p_{L|D,SH}$ and $p_{L|D,SL}$ because the probability of transitions from S-states to D-states is zero. We encode the requirement that "decisions to start a green transition, if implemented, are more likely to yield states with a low level of economic wealth (L-states) than states with high economic wealth" by the specification

$$\mathit{pSpec3} : p_{H|S,DH} \leqslant p_{L|S,DH}$$

Because $p_{H|S,DH} = 1 - p_{L|S,DH}$, this requires $p_{L|S,DH}$ to be greater or equal to 50%. Let's say that

$$p_{L|S,DH} = \ 0.7$$

We also want to express the idea that starting a green transition in a weak economy (perhaps a sub-optimal decision?) is more likely to yield a weak economy than starting a green transition in a strong economy

$$\mathit{pSpec4} : p_{L|S,DH} \leqslant p_{L|S,DL}$$

which requires specifying a value of $p_{L|S,DL}$ between 0.7 and 1.0, say

$$p_{L|S,DL} = 0.9$$

This fixes the values of $p_{L|S,DH}$ and $p_{L|S,DL}$ for our decision process in the ranges imposed by the "semantic" constraints *pSpec3* and *pSpec4*. We discuss how these (and other) transition probabilities would have to be estimated in a more realistic (as opposed to stylised) GHG emissions decision process in section 6.

Next, we have to specify the remaining transition probabilities $p_{L|S,SH}$, $p_{L|S,SL}$, $p_{L|D,DH}$ and $p_{L|D,DL}$. What are meaningful constraints for these? Remember that $p_{L|S,SH}$ and $p_{L|S,SL}$ represent the probabilities of transitions to low wealth states (L-states) from H and L-states, respectively, while an already started green transition is accomplished. In this situation, and again because of the inertia of economic systems, it is reasonable to assume that transitions from H-states (booming economy) to H-states are more likely than transitions from H-states to L-states and, of course, the other way round. In formulas:

$$\mathit{pSpec5} : p_{L|S,SH} \leqslant p_{H|S,SH}$$
$$\mathit{pSpec6} : p_{H|S,SL} \leqslant p_{L|S,SL}$$

Again, because $p_{H|S,SH} = 1 - p_{L|S,SH}$ (and $p_{H|S,SL} = 1 - p_{L|S,SL}$), this requires $p_{L|S,SH}$ and $p_{L|S,SL}$ to be below and above 50%, respectively.

In our decision process, a high value of $p_{L|S,SL}$ implies a low probability of recovering from economic downturns in states in which a transition towards a globally decarbonised society has been started or has been accomplished.

In more realistic specifications of GHG emission processes, one may want to distinguish between these two cases, or even to keep track of the time elapsed since a green transition was started and define the probability of recovering from economic downturns accordingly.

Conversely, a low value of $p_{L|S,SH}$ means high *resilience* against economic downturns in states in which a transition towards a globally decarbonised society has been started or has been accomplished. In such states, we assume a moderate likelihood of fast recovering from economic downturns:

$$p_{L|S,SL} = 0.7$$

and also a moderate resilience

$$p_{L|S,SH} = 0.3$$

Let's turn the attention to the last two transition probabilities that need to be specified in order to complete the description of the transitions leading to economic downturns or recoveries. These are $p_{L|D,DH}$ and $p_{L|D,DL}$.

The semantics of $p_{L|D,DH}$ and $p_{L|D,DL}$ should meanwhile be clear: $p_{L|D,DH}$ represents the probability of economic downturns and $1 - p_{L|D,DL}$ the probability of recovering (from economic downturns) in states in which a green transition has not already been started. As for their counterparts discussed above, we have the semantic requirements

$$\mathit{pSpec7} : p_{L|D,DH} \leqslant p_{H|D,DH}$$
$$\mathit{pSpec8} : p_{H|D,DL} \leqslant p_{L|D,DL}$$

with $p_{H|D,DH} = 1 - p_{L|D,DH}$ and $p_{H|D,DL} = 1 - p_{L|D,DL}$ and thus, by the same argument as for $p_{L|S,SH}$ and $p_{L|S,SL}$, $p_{L|D,DH}$ and $p_{L|D,DL}$ below and above 50%, respectively.

How should $p_{L|D,DH}$ and $p_{L|D,DL}$ compare to $p_{L|S,SH}$ and $p_{L|S,SL}$? Is the likelihood of economic downturns in states in which a green transition has not already been started higher or lower than the likelihood of economic downturns in states in which a transition towards a globally decarbonised society has been started or has been accomplished? Realistic answers to this question are likely to depend on the decision step and on the time elapsed since the green transition has been started, see 6. As a first approximation, here we just assume that these probabilities are the same:

$$\begin{aligned} p_{L|D,DL} &= p_{L|S,SL} \\ p_{L|D,DH} &= p_{L|S,SH} \end{aligned}$$

This completes the discussion of the probabilities of economic downturns and recoveries.

- *The probabilities of commitment to severe impacts from climate change:*

  The last ingredient that we need to fully specify the transition function of our decision process are the probabilities of transitions to states that are committed to severe impacts from climate change. In the introduction, we have stipulated that

  > ... we assume that the probability of entering states in which the world is committed to future severe impacts from climate change is higher in states in which a green transition has not already been started as compared to states in which a green transition has been started.

  We account for this assumption with four transition probabilities: $p_{U|S,0}$, $p_{U|D,0}$, $p_{U|S}$ and $p_{U|D}$. The first two represent the probabilities of transitions (from uncommitted states) to uncommitted states at decision step zero for the cases in which a transitions to a decarbonised economy has been implemented and delayed, respectively. Similarly, $p_{U|S}$ and $p_{U|D}$ represent the probabilities of transitions from U-states to U-states at later decision steps. We take the informal specification

  > ... delaying transitions to decarbonised economies increases the likelihood of entering states in which the world is committed to future severe impacts from climate change.

  by the letter and, for the sake of simplicity, assume that the whole increase in the likelihood of entering committed states takes place during the first step of our decision process. This is a very crude assumption and we will come back to it when we discuss the results of measures of responsibility in section 7. With these premises (and keeping in mind that $p_{C|S,0} = 1 - p_{U|S,0}$, $p_{C|D,0} = 1 - p_{U|D,0}$, etc.) our informal specification translates into the constraints:

  $$\begin{aligned} &\mathit{pSpec9} &&: p_{C|S,0} \leqslant p_{U|S,0} \\ &\mathit{pSpec10} &&: p_{C|S,0} \leqslant p_{C|D,0} \\ &\mathit{pSpec11} &&: p_{C|S} \quad \leqslant p_{U|S} \\ &\mathit{pSpec12} &&: p_{C|S} \quad \leqslant p_{C|D} \\ &\mathit{pSpec13} &&: p_{C|D,0} \leqslant p_{C|D} \end{aligned}$$

  For the time being, we set $p_{U|S,0}$, $p_{U|D,0}$, $p_{U|S}$ and $p_{U|D,0}$ to 0.9, 0.7, 0.9 and 0.3, respectively. In words, we assume a 30% chance of committing to future severe impacts from climate change if we fail to start a green transition at the first decision step. We assume this chance to increase to 70% at later decision steps. We also assume a 10% chance of severe climate change impacts if we start a green transition at the first decision step or later.

With the transition probabilities in place, we can now specify the transition function of the decision process. For illustration, we first discuss the definition of the transition function for one specific case, before we give a more compact definition based on a Bayesian network.

Recall from Section 2 that the transition function gets as inputs the decision step, the current state and the current control/decision. Consider the case at step zero in which the (initial) state is (D, H, U) and the control is *Start*, i.e. the decision is to start a green transition:

$$\mathit{Theory.next}\ Z\ (D, H, U)\ \mathit{Start} = \mathit{mkSimpleProb}$$

$$
\begin{aligned}
[&((D, H, U), p_{D \mid Start} * p_{H \mid D,DH} * p_{U \mid D,0}),\\
&((D, H, C), p_{D \mid Start} * p_{H \mid D,DH} * p_{C \mid D,0}),\\
&((D, L, U),\ p_{D \mid Start} * p_{L \mid D,DH}\ * p_{U \mid D,0}),\\
&((D, L, C),\ p_{D \mid Start} * p_{L \mid D,DH}\ * p_{C \mid D,0}),\\
&((S, H, U),\ p_{S \mid Start}\ * p_{H \mid S,DH}\ * p_{U \mid S,0}),\\
&((S, H, C),\ p_{S \mid Start}\ * p_{H \mid S,DH}\ * p_{C \mid S,0}),\\
&((S, L, U),\ \ p_{S \mid Start}\ * p_{L \mid S,DH}\ \ * p_{U \mid S,0}),\\
&((S, L, C),\ \ p_{S \mid Start}\ * p_{L \mid S,DH}\ \ * p_{C \mid S,0})]
\end{aligned}
$$

The result of the transition is a finite probability distribution on possible next states in which the probabilities of reaching the respective states are determined in a compositional manner from the transition probabilities introduced in the previous subsection. We only comment the definition of the probability of (S, H, U), the state in which a green transition has been started, the economy is in a wealthy state and the world is not committed to future severe impacts from climate change. The other states' probabilities can be interpreted analogously.

The probability of (S, H, U) is defined as the product of three transition probabilities: the probability that a green transition is actually implemented, given that the decision was to do so $p_{S \mid Start}$; the probability that the economy is in a good state (an H-state) given that a green transition has been started from an H-state $p_{H \mid S,DH}$: the probability of entering states that are not committed to severe impacts from climate change, again given that a transitions to a decarbonised economy has been started $p_{U \mid S,0}$.

Notice that $p_{C \mid D,0} + p_{U \mid D,0}$ and $p_{C \mid S,0} + p_{U \mid S,0}$ are equal to one by definition of $p_{C \mid D,0}$ and $p_{C \mid S,0}$. The same holds for $p_{H \mid D,DH} + p_{L \mid D,DH}$ and $p_{H \mid S,DH} + p_{L \mid S,DH}$ (by definition of $p_{H \mid D,DH}$, $p_{H \mid S,DH}$) and for $p_{D \mid Start} + p_{S \mid Start}$ (by definition of $p_{D \mid Start}$). It follows that the sum of the probabilities of *next Z DHU Start* is one, as one would expect.

It is possible to derive the probability of (S, H, U) (and of all other possible next states) given the decision to *Start* a green transition in (D, H, U)

$$p_{S \mid Start} * p_{H \mid S,DH} * p_{U \mid S,0}$$

rigorously if we represent our stylised decision process as a Bayesian belief network. To this end, it is useful to introduce some notation from elementary probability theory. Different textbooks adopt slightly different notations; here, we follow [Mit97] and denote the conditional probability of entering (S, H, U) given the decision to *Start* a green transition in (D, H, U) with

$$P\ (S, H, U \ \mid\ Start, D, H, U)$$

Thus, our obligation is to show

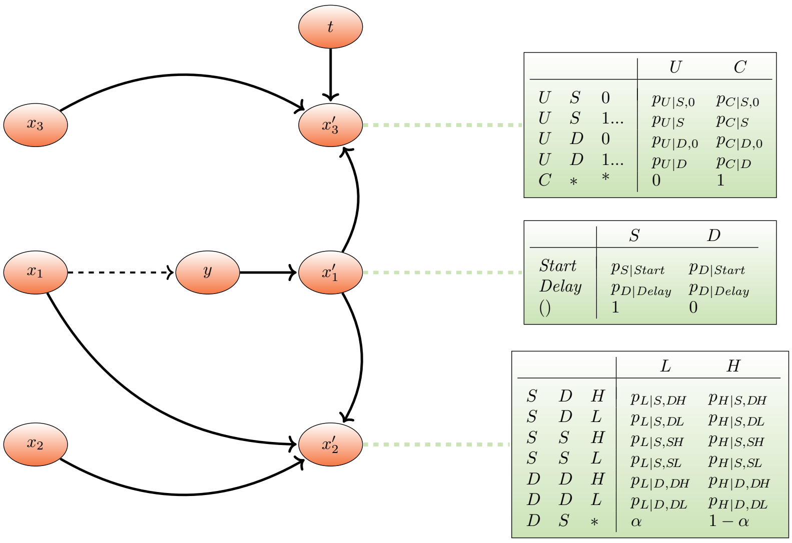
$$P\ (S, H, U \ \mid\ Start, D, H, U) = p_{S \mid Start} * p_{H \mid S,DH} * p_{U \mid S,0}$$

Let x_1, $x2$, $x3$ denote the components of the *current* state $x : X\ t$ and $x1'$, $x2'$, $x3'$ the components of the *next* state. Thus, for $x = (D, H, U)$, we have $x_1 = D$, $x2 = H$ and $x3 = U$. As usual, we denote a decision in x at step t with $y : Y\ t\ x$.

The *variables* x_1, $x2$, $x3$, y, $x1'$, $x2'$, $x3'$ and the decision step t are associated with the *nodes* of the Bayesian network of figure 3. The *edges* of the network encode the notion of conditional dependency: the arrow between x_1 and $x2'$ posits that the probability of transitions to states with a low (high) economic wealth depends on whether a green transition is currently underway or has been delayed[11].

The conditional probability tables associated with the nodes, encode such probabilities. Thus, for example, the table associated with $x1'$ posits that the conditional probability of entering S-states given that the decision (variable y) was to *Start* a green transition is $p_{S \mid Start}$ as discussed above. Similarly, the table associated with $x2'$ encodes the specification that the probability of entering an

[11] Because of the arrows between $x2$ and $x1'$ and $x2'$, such probability also depends on whether the current state of the economy is low or high and on whether a green transition gets started or not.



			U	C
U	S	0	$p_{U\|S,0}$	$p_{C\|S,0}$
U	S	1...	$p_{U\|S}$	$p_{C\|S}$
U	D	0	$p_{U\|D,0}$	$p_{C\|D,0}$
U	D	1...	$p_{U\|D}$	$p_{C\|D}$
C	$*$	*	0	1

	S	D
Start	$p_{S\|Start}$	$p_{D\|Start}$
Delay	$p_{D\|Delay}$	$p_{D\|Delay}$
()	1	0

			L	H
S	D	H	$p_{L\|S,DH}$	$p_{H\|S,DH}$
S	D	L	$p_{L\|S,DL}$	$p_{H\|S,DL}$
S	S	H	$p_{L\|S,SH}$	$p_{H\|S,SH}$
S	S	L	$p_{L\|S,SL}$	$p_{H\|S,SL}$
D	D	H	$p_{L\|D,DH}$	$p_{H\|D,DH}$
D	D	L	$p_{L\|D,DL}$	$p_{H\|D,DL}$
D	S	$*$	α	$1-\alpha$

Figure 3: Stylised decision process as a Bayesian network.

L-state given that an S-state was entered from a current D- and H-state is $p_{L|S,DH}$[12].

With a Bayesian network representation of our stylised decision process in place, we can derive

$$P\ (S, H, U\ \mid\ Start, D, H, U) = p_{S|Start} * p_{H|S,DH} * p_{U|S,0}$$

rigorously by equational reasoning. The computation is straightforward but we spell out each single step for clarity:

$P\ (S, H, U\ \mid\ Start, D, H, U)$

$=$ -- definition of $x1'$ ... y ... $x3$

$P\ (x1' = S, x2' = H, x3' = U\ \mid\ y = Start, x_1 = D, x2 = H, x3 = U)$

$=$ -- definition of conditional probability, set theory

$P\ (x2' = H, x3' = U, x1' = S\ \mid\ y = Start, x_1 = D, x2 = H, x3 = U)$

$=$ -- chain rule

$P\ (x2' = H\ \mid\ x3' = U, x1' = S, y = Start, x_1 = D, x2 = H, x3 = U) *$
$P\ (x3' = U, x1' = S\ \mid\ y = Start, x_1 = D, x2 = H, x3 = U)$

$=$ -- chain rule

$P\ (x2' = H\ \mid\ x3' = U, x1' = S, y = Start, x_1 = D, x2 = H, x3 = U) *$
$P\ (x3' = U\ \mid\ x1' = S, y = Start, x_1 = D, x2 = H, x3 = U) *$
$P\ (x1' = S\ \mid\ y = Start, x_1 = D, x2 = H, x3 = U)$

$=$ -- Bayesian network (conditional independence)

$P\ (x2' = H\ \mid\ x1' = S, x_1 = D, x2 = H) * P\ (x3' = U\ \mid\ x1' = S, x3 = U) * P\ (x1' = S\ \mid\ y = Start)$

$=$ -- Bayesian network (tables)

$p_{H|S,DH} * p_{U|S,0} * p_{S|Start}$

Similar derivations can be obtained, in terms of the network of Fig. 3, for the other transition probabilities that define *next* $Z\ (D, H, U)$ *Start* and, in fact, for all the transition probabilities that define *next*. Thus, Fig. 3 can be seen as a compact representation of the transition function *next* of our stylised decision process.[13] The conditional probability tables can be translated into probability

[12] Notice that the conditional probability table associated with $x2'$ contains an undefined value α. This is because the probability of entering L (or H) states given that a D-state was entered starting from an S-state is irrelevant: the probability of transitions from S-states to D-states is zero (remember that we have assumed that green transitions cannot be halted or reversed by later decisions), as encoded in the third row of the table associated with $x1'$.

[13] Notice that the causal networks at the core of the storyline approach [She19] are also Bayesian belief networks,

functions in a straightforward way and then be used to define the transition function *next* without having to spell out all individual cases.

We represent the three probability tables from Fig. 3 as follows:

- the probability of transitioning into an S/D-state if the decision is *Start*/ *Delay*

$$\begin{array}{llll}
\mathit{condX1'} : \{t : \mathbb{N}\} \to \{x : X\ t\} \to Y\ t\ x \to \mathit{SimpleProb}\ \mathit{SD} \\
\mathit{condX1'}\ \{x = (D, \_x2, \_x3)\}\ \mathit{Start} & = \mathit{mkSimpleProb}\ [(S, p_{S|Start}), & (D, p_{D|Start})] \\
\mathit{condX1'}\ \{x = (D, \_x2, \_x3)\}\ \mathit{Delay} & = \mathit{mkSimpleProb}\ [(S, p_{S|Delay}), & (D, p_{D|Delay})] \\
\mathit{condX1'}\ \{x = (S, \_x2, \_x3)\}\ () & = \mathit{mkSimpleProb}\ [(S,\ 1.0), & (D,\ 0.0)]
\end{array}$$

- the probability of transitioning to an L/H-state, if the next start is an S/D-state and the current state an S/D-state with with L/H economic wealth (recall from above that α is a placeholder value for an impossible case):

$$\begin{array}{lll}
\mathit{condX2'} : \mathit{SD} \to \mathit{SD} \to \mathit{LH} \to \mathit{SimpleProb}\ \mathit{LH} \\
\mathit{condX2'}\ S\ D\ H & = \mathit{mkSimpleProb}\ [(L, p_{L|S,DH}), & (H, p_{H|S,DH})] \\
\mathit{condX2'}\ S\ D\ L & = \mathit{mkSimpleProb}\ [(L, p_{L|S,DL}), & (H, p_{H|S,DL})] \\
\mathit{condX2'}\ S\ S\ H & = \mathit{mkSimpleProb}\ [(L, p_{L|S,SH}), & (H, p_{H|S,SH})] \\
\mathit{condX2'}\ S\ S\ L & = \mathit{mkSimpleProb}\ [(L, p_{L|S,SL}), & (H, p_{H|S,SL})] \\
\mathit{condX2'}\ D\ D\ H & = \mathit{mkSimpleProb}\ [(L, p_{L|D,DH}), & (H, p_{H|D,DH})] \\
\mathit{condX2'}\ D\ D\ L & = \mathit{mkSimpleProb}\ [(L, p_{L|D,DL}), & (H, p_{H|D,DL})] \\
\mathit{condX2'}\ D\ S\ \_x2 & = \mathit{mkSimpleProb}\ [(L,\ \alpha), & (H,\ (1.0 - \alpha))]
\end{array}$$

- the probability of transitioning to an U/C-state from an S/D-state at step $n : \mathbb{N}$:

$$\begin{array}{lll}
\mathit{condX3'} : \mathit{UC} \to \mathit{SD} \to \mathbb{N} \to \mathit{SimpleProb}\ \mathit{UC} \\
\mathit{condX3'}\ U\ S\ \ Z & = \mathit{mkSimpleProb}\ [(U, p_{U|S,0}), & (C, p_{C|S,0})] \\
\mathit{condX3'}\ U\ S\ \ \_n & = \mathit{mkSimpleProb}\ [(U, p_{U|S}), & (C, p_{C|S})\ ] \\
\mathit{condX3'}\ U\ D\ \ Z & = \mathit{mkSimpleProb}\ [(U, p_{U|D,0}), & (C, p_{C|D,0})] \\
\mathit{condX3'}\ U\ D\ \ \_n & = \mathit{mkSimpleProb}\ [(U, p_{U|D}), & (C, p_{C|D})] \\
\mathit{condX3'}\ C\ \_x1\ \_n & = \mathit{mkSimpleProb}\ [(U,\ 0.0), & (C,\ 1.0)]
\end{array}$$

Based on *condX1'*, *condX2'* and *condX3'*, we can now compute the conditional probability of a possible next state.

$$\begin{array}{l}
\mathit{jointCondProb} : (t : \mathbb{N}) \to (x : X\ t) \to (y : Y\ t\ x) \to (x : X\ (S\ t)) \to \mathit{Double}_{\geqslant 0} \\
\mathit{jointCondProb}\ t\ (x_1, x2, x3)\ y\ (x1', x2', x3') = \\
\quad \mathit{prob}\ (\mathit{condX1'}\ y)\ x1' * \mathit{prob}\ (\mathit{condX2'}\ x1'\ x_1\ x2)\ x2' * \mathit{prob}\ (\mathit{condX3'}\ x3\ x1'\ t)\ x3'
\end{array}$$

where the function

$$\mathit{prob} : \mathit{Eq}\ A \Rightarrow \mathit{SimpleProb}\ A \to A \to \mathit{Double}_{\geqslant 0}$$

returns the probability of an outcome $a : A$ according to a probability distribution $\mathit{spa} : \mathit{SimpleProb}\ A$. Thus, if we return to our example from above with $(x_1, x2, x3) = (D, H, U)$ and $(x1', x2', x3') = (S, H, U)$

$$\mathit{prob}\ (\mathit{condX1'}\ \mathit{Start})\ S$$

is the conditional probability $P(x_1' = S \mid y = \mathit{Start}) = \mathit{pS\_Start}$,

$$\mathit{prob}\ (\mathit{condX2'}\ S\ D\ H)\ H$$

is $P(x_2' = H \mid x_1' = S, x_1 = D, x_2 = H) = \mathit{pH\_S\_DH}$ and

$$\mathit{prob}\ (\mathit{condX3'}\ U\ S\ 0)\ U$$

is $P(x_3' = U \mid x_3 = U, x_1' = S, t = 0) = \mathit{pU\_S\_}0$.

albeit without a clear-cut distinction between state and control spaces.

Collecting all of our states into a list

```
states : List State
states = [(x₁, x2, x3) | x₁ ← [S, D], x2 ← [L, H], x3 ← [U, C]]
```

we can then define the transition function uniformly for all inputs using list comprehension:

```
next : (t : ℕ) → (x : X t) → Y t x → M (X (S t))


next t x y =
  mkSimpleProb {prf = prf′ t x y}
    [(x′, p) | x′ ← states,
               p  ← [jointCondProb t x y x′],
               0.0 < toDouble p]
```

In standard mathematical notation, writing *jointCondProb t x y x′* as $\pi_{t,x,y,x'}$, the list comprehension in the definition of *next t x y* can be understood like a set comprehension defining a set of state-probability pairs

$$SP_{t,x,y} = \{(x', p) \mid x' \in State, p = \pi_{t,x,y,x'}, p > 0\}$$

The condition $p > 0$ ensures that states with zero probability – i.e. *im*possible next states – are not included.

6 Interlude: Realistic and stylised processes

Before continuing the technical part of the document, let us clarify the notion of *stylised* decision process. As mentioned in the introduction, this notion was originally introduced in [BJI18] to contrast the one of *realistic* decision process. This is also the sense in which it has been used in this report.

For example, in discussing the probability of economic downturns, we have argued that, in the specification of more realistic GHG emissions decision processes, one might want to distinguish between states in which a transition towards a globally decarbonised society is ongoing and states in which the transition has already been accomplished. In the case of ongoing green transitions, one may want to consider different transition probabilities, perhaps depending on the degree to which the transition has been accomplished or the time since it was started.

From this angle, more realistic essentially means a larger number of states (remember that, as discussed in the introduction, the states of a decision process typically represent sets of micro-states with the latter being detailed descriptions of physical, economic and social conditions), perhaps also of control options (for example, fast or slow green transitions) and hence more complex transition functions.

This reductionist approach towards "realism" is paradigmatic of so-called *modelling* approaches. In climate policy advice, it has lead to (integrated assessment) models of decision processes based on high-dimensional state and control spaces and a large number of model parameters [Nor18, HWFD19].

While this is popular in climate policy assessment and advice, the usage of "realistic" integrated assessment models (IAM) has also been criticised, among others, because of their poor understandability and limited predictive capability. For example, in [Pin17], it was found that very different estimates of the "right" social cost of carbon can be "obtained" by setting the values of certain IAM parameters (for example, discount factors and climate sensitivities) to specific, arbitrary but "plausible" values and Pindyck even argued that

> IAM-based analyses of climate policy create a perception of knowledge and precision that is illusory and can fool policymakers into thinking that the forecasts the models generate have some kind of scientific legitimacy [Pin17].

Similar concerns and the problem that a too strong focus on *reliability* may be unsuitable for climate decision making at regional scales, have been discussed in [She19].

Another weakness of IAMs for climate policy is their strong bias towards deterministic modelling. With very few exceptions, these models assume that decisions (e.g. of starting a global green transition) are implemented with certainty, that crucial parameterizations of climate processes (like the *equilibrium climate sensitivity*) can be estimated accurately and that the costs and the benefits of future climate changes can be accounted for in suitable "terminal" (salvage, scrap, see [Put14] section 2.1.3) rewards.

Is there a way of specifying decision processes that are useful for pragmatic climate decision making and that avoid the drawbacks of deterministic modelling approaches based on high-dimensional state spaces?

We believe that this is the case and that, rather than neglecting uncertainty, the way to address this challenge is to 0) specify low-dimensional state and control spaces that are logically consistent with the informal description of the specific decision process at stake; 1) explicitly account for the uncertainties that are known to affect best decisions for that process, 2) exploit the knowledge available (from past experience, data, model simulation, etc.) to specify trustable transition probabilities with interpretations that are consistent with that process.

This is the essence of the approach that we have demonstrated in the previous section: starting from an informal description, we have introduced formal specifications of state and control spaces that are logically consistent with that description. We have accounted for the uncertainties of the informal description in terms of twelve transition probability parameters. For each parameter, we have provided an interpretation together with a range of values compatible with that interpretation. Within these ranges, we have then chosen certain values and defined the transition function in terms of those values. For example, we have postulated a 10% chance that a decision to start a green transition fails to be implemented.

In a (more) realistic specification, this figure could perhaps have been obtained by asking a pool of experts, perhaps political scientists, historians, etc. Similarly, in more realistic specifications, the probabilities of recovering from economic downturns might be obtained from climate economists. These, in turn, might rely on model simulations, expert elicitation or perhaps statistical data. Finally, climate models (general circulation models, intermediate complexity models, low-dimensional systems of ordinary differential equations representing global mass and energy budgets) might be applied to representative micro-states samples of a given (macro) state (for example, our initial state *DHU*) to compute more realistic estimates (for example via Monte Carlo simulations) of transition probabilities, for instance, to committed states.

From this angle, the approach of "stylised" decision processes is similar to the *storyline* approach – the "identification of physically self-consistent, plausible pathways" – proposed in [She19]. The focus, there on physical consistency and causal networks, is here on logical consistency and decision networks. Common to both approaches is the need to integrate contributions from very different disciplines, ranging from theoretical computer science to the social sciences [She19, SMV^{+}21].

In this enterprise, the theory of section 2 and the language extensions discussed in this report play a twofold role. On the one hand, they help ensuring that results of model simulations, expert opinions, and statistical data are applied consistently. On the other hand, they make it possible to reason about pragmatic decision processes in a formal and rigorous way.

7 Generic responsibility measures

In [BBC^{+}21] we introduced generic responsibility measures to answer the following questions:

- What does it mean precisely for decisions to matter?
- Are there general ways to measure how much decisions matter when these have to be taken under uncertainty?
- Is there a natural way of comparing similar decisions at different times?

To define a notion of responsibility we use and extend the basic theory for monadic SDPs as follows:

(S1) The *reward* function of the SDP is defined based on a specific *goal* of decision making. For example, a formal expression of *"avoiding states committed to severe climate change impacts"*.

(S2) Verified "best" and "conditional worst" decisions are compared at the specific state at which we want to measure how much decisions matter for the goal encoded in (S1).

(S3) We define a degree of responsibility consistent with this measure.

In [BvH18] three conditions are put forward under which "a person can be ascribed responsibility for a given outcome":

(C1) *avoidance*: it is possible for the person to avoid an performing an action that contributes to the outcome,

(C2) *agency*: having the capability to act intentionally, to plan, and to distinguish between desirable and undesirable outcomes, and

(C3) *causal relevance*: there is a causal relation between the person's action and the outcome.

The notion of causality is not uncontroversial [Car95] and its role in formalizations of responsibility has been addressed, among others by [CH04, Hal06] and [Hal14]. Below we show that at least for sequential decision processes it is possible to define "meaningful" measures of how much decisions matter without having to deal with causality. We also discuss the relation between these measures and responsibility measures.

7.1 Illustration of the approach

For concreteness, we illustrate (S1)-(S3) for the decision problem of section 5. The extensions of the theory discussed in this section, however, are fully generic and can be applied to arbitrary decision processes. We tackle step one by first discussing conditions under which decisions shall not matter.

(S1)a: When decisions shall not matter. Consider the problem of attributing a non-negative number to the states of a decision process P:

$$\mathit{mMeas} : (t : \mathbb{N}) \to X\ t \to \mathit{Double}_{\geqslant 0}$$

The idea is that $\mathit{mMeas}\ P\ t\ x$ represents how much decisions in state x do matter: the larger, the more decisions in x matter. For the time being, assume that $\mathit{mMeas}\ t\ x$ takes values between zero and one. Under which conditions shall we require it to be zero? First and foremost we would like $\mathit{mMeas}\ t\ x$ to be zero whenever only one option is available to the decision maker in x:

$$\mathit{mMeasSpec1} : (t : \mathbb{N}) \to (x : X\ t) \to \mathit{Singleton}\ (Y\ t\ x) \to \mathit{mMeas}\ t\ x = \mathit{zero}$$

Here, we have formalised the condition that only one option is available to the decision maker in x with the predicate $\mathit{Singleton}\ (\mathit{Ctrl}\ P\ t\ x)$.

The specification *mMeasSpec1* is consistent with *avoidance*, one of the three conditions of [BvH18] listed above.

(S1)b: Encoding goals of decision making. To measure how much decisions matter, we have to extend a decision process to a decision problem by providing the definitions of *Val*, *reward*, *meas*, $\oplus$, $\sqsubseteq$ and *zero*. In the following, we define these components for our stylised decision process of Section 5. In our example, we have a stochastic SDP for which the type and structure used for valuation are simply

$$\begin{aligned} \mathit{Val}\ &= \mathit{Double}_{\geqslant 0} \\ (\oplus)\ &= (+) \end{aligned}$$

$$zero = 0.0$$
$$(\sqsubseteq) = (\leqslant)$$

Here, we follow standard decision theory and take *meas* to be the expected value measure, although other measures might be used as long as they fulfil the compatibility conditions stated in Section 2.

$$meas = expectedValue$$

As a starting point for the definition of the reward function, we identify the *goal* for which we seek responsibility measures. In short, we have to say *for what* we want to measure how much decisions matter. For example, we might be interested in measuring how much decisions matter for *avoiding states that are committed to severe impacts from climate change.* Or perhaps we want to measure how much decisions matter for *avoiding climate change impacts but also economic downturns.* Formally, we may express these goals using the test functions *isCommitted* and *isDisrupted* which we defined in Section 5.

$$goal : \{t : \mathbb{N}\} \to (X\ t) \to \mathbb{B}$$

with

$$goal\ \{t\}\ x = isCommitted\ t\ x$$

or

$$goal\ \{t\}\ x = isCommitted\ t\ x \vee isDisrupted\ t\ x$$

Then we can define the reward function in terms of this goal:

$$reward\ t\ x\ y\ x' = \mathbf{if}\ goal\ \{t = S\ t\}\ x'\ \mathbf{then}\ \ 0.0\ \mathbf{else}\ \ 1.0$$

In the following, we will use the second definition of *goal*. In Section 7.2 below, we discuss generic goal functions and show how to pre-define *Val*, *reward*, etc. based on such functions.

(S2): Measuring how much decisions matter. With a goal encoded via the reward function, we can tackle the problem of measuring how much decisions in a state do matter for that goal.

For concreteness, let's consider the initial state (D, H, U) of our decision problem. In this state, the decision maker has two options: start a green transition or further delay it. Remember that our decision maker is effective only to a certain extent. As shown in figure 3, a decision to start a green transition may well yield a next state in which the transition has been delayed. According to Section 5, the probability of this event is $p_{D|Start}$, that is, 10%.

What does this uncertainty imply for the decision to be taken in the initial state (D, H, U)? Answering this question rigorously requires fixing a decision *horizon.* This is the number of decision steps of our decision process that we look ahead in order to measure how much decisions matter. Remember from section 2 that the value of taking zero decision steps is always $zero : Val$, a problem-specific reference value that holds for every decision step and state at that step. Thus, if we look forward zero steps, no decision matters, independently of the decision step and state. But, for a strictly positive number of decision steps, we can formulate and rigorously answer the following two questions

- Is it better, in (D, H, U) to (decide to) start or to delay a green transition?
- How much does this decision matter (for avoiding climate change impacts but also economic downturns)?

To do so, we first apply the generic backward induction from Section 2 and compute an optimal sequence of policies *ps* over the horizon. Recall from Sectionsection:valval that *bi* computes provably optimal policy sequences[14]. This means that no other policy sequence entails better decisions than *ps* (here for the goal of avoiding climate change impacts but also economic downturns).

[14] If $\sqsubseteq$, $\oplus$, *meas*, etc. fulfil the specifications from Section 2.

Thus, we can compute a best decision and the value (of the sum of the rewards) over a horizon of n steps for arbitrary states:

```
best :(t, n : ℕ) → X t → String
best t   Z    x = "The horizon must be greater than zero!"
best t ( S m) x =
  let ps = bi (S t) m in
  let p  = bestExt ps in
  let b  = p x in
  let vb = val (p :: ps) x in
  "Horizon, best, value:  " ++
  show (S m) ++ ",  " ++
  showY b ++ ",  " ++
  show vb
```

What is a best decision in (D, H, U) for a horizon of only one step?

```
* Responsibility > : exec best 0 1 (D, H, U)
Horizon, best, value : 1, Delay, 0.468
```

This is not very surprising: according to the definition of *next*, the probability of entering states that are either economically disrupted or committed to severe impacts from climate change is 0.708. Thus, the expected value of deciding to start a green transition is only

$$1 - 0.708 = 0.292$$

By contrast, the expected value of deciding to delay a green transition is 0.468, as seen above. As it turns out, one has to look forward at least over three decision steps (or, in our interpretation, about three decades) for the decision to start a green transition to become a best decision in (D, H, U). We can apply the computation

```
bests :(t : ℕ) → List ℕ → X t → IO ()
bests t Nil       x = putStrLn "done!"
bests t (n :: ns) x = do putStrLn (best t n x)
                         bests t ns x
```

to study how best decisions vary with the horizon. Again, for $x = (D, H, U)$ one obtains:

```
* Responsibility > : exec bests 0 [1 .. 8] (D, H, U)
Horizon, best, value : 1, Delay, 0.468
Horizon, best, value : 2, Delay, 0.635454
Horizon, best, value : 3, Start, 0.940669612
Horizon, best, value : 4, Start, 1.250012318344
Horizon, best, value : 5, Start, 1.533635393558128
Horizon, best, value : 6, Start, 1.790773853744118
Horizon, best, value : 7, Start, 2.022874449805313
```

As anticipated, the decision to start a green transition at the first decision step becomes a best decision for horizons of three or more decisions. The other way round: our decision maker would have to be very myopic (or, equivalently very much discount future benefits) to conclude that delaying a green transition is a best decision in (D, H, U).

But how much does this decision actually matter? To answer this question, we need to compare a best decision in (D, H, U) *for a given time horizon* to a worst decision. Again, for concreteness, let's for the moment fix the horizon to 7 decision steps.

What is the value (again, in terms of the sum of the rewards associated with avoiding climate change impacts and economic downturns) of deciding to delay a green transition in (D, H, U)? There are different ways of answering this question, but a canonical one[15] is to consider the consequences of deciding to delay a green transition at the first decision step in (D, H, U) *and* take later decisions

[15] We discuss alternative approaches in section 7.3.

optimally. In our specific problem, this corresponds to assuming that future generations will do their best to avoid negative impacts from climate change and economic downturns.

If we denote our optimal policy sequence for an horizon of 7 steps by *ps*, we can compute the consequences of deciding to delay at the first decision step in (D, H, U) and then take later decisions optimally by replacing the first policy of *ps* with one that recommends *Delay* in (D, H, U):

```
ps  :  PolicySeq 0 7
ps = bi 0 7

ps′ : PolicySeq 0 7
ps′ = (setInTo (head ps) (D, H, U) Delay) :: tail ps
```

The function *setInTo* in the definition of ps' is a higher-order primitive: it takes a function (in this case the first policy of *ps*), a value in its domain and one in its codomain, and returns a function of the same type that fulfils the specification

$$(setInTo\ f\ \ a\ \ b)\ a = b\ \ \wedge\ \ Not\ (a = a') \to (setInTo\ f\ \ a\ \ b)\ a' = f\ \ a'$$

for all f, a, a' and b of appropriate type. With ps', we can compute the value of deciding to delay a green transition at the first decision step in (D, H, U):

```
∗ Responsibility > : exec show [val ps (D, H, U), val ps′ (D, H, U)]
"[2.022874449805313, 1.672795254555656]"
```

The difference between the value of *ps* and the value of ps' in (D, H, U) then *is a measure of how much decisions in (D, H, U) matter for avoiding climate change impacts and economic downturns over a time horizon of 7 decision steps*: the bigger this difference, the more the decision matters.

S3: Responsibility measures We have argued that the difference between the value of *ps* and the value of ps' in (D, H, U), is a measure of how much decisions in (D, H, U) matter for avoiding climate change impacts and economic downturns over a time horizon of 7 decision steps. This argument is justified because:

- We have defined optimal policy sequences to be policy sequences that avoid (as well as it gets) climate change impacts and economic downturns in (S1).
- Over 7 decision steps, *ps* is a *verified optimal* policy sequence.
- The best decision in (D, H, U) is to start a green transition:

  ```
  ∗ Responsibility > : exec show (head ps (D, H, U))
  "Start"
  ```

- ps' is a sequence of policies identical to *ps* except for recommending *Delay* instead of *Start* in (D, H, U) and for the first decision step:

  ```
  ∗ Responsibility > : exec show (head ps′ (D, H, U))
  "Delay"
  ```

These facts are sufficient to guarantee that the difference between the value of *ps* and the value of ps' in (D, H, U) is actually the difference between the value (with respect to *goal*) of the best and of the worst decisions that can be taken in (D, H, U).

The computation and the definitions of *ps* and ps' suggest a refinement and an implementation of the measure of how much decisions matter *mMeas* put forward in the beginning of this section. First, we want *mMeas* to depend on a time horizon n. Second, for the sake of simplicity, we want *mMeas* to return plain double precision floating point numbers:

```
mMeas :(t : ℕ) → (n : ℕ) → X t → Double
mMeas t Z x       = 0.0
```

```
mMeas t (S m) x = let ps = bi (S t) m in
                  let v  = toDouble (val (bestExt ps :: ps) x) in
                  let v' = toDouble (val (worstExt ps :: ps) x) in
                  v − v'
```

Remember that, in (S1), we have encoded the goal of avoiding severe climate change impacts and economic downturns for which we compute *mMeas* through a function

```
reward t x y : X (S t) → Double⩾0
```

that returns 0 for next states that are committed to severe climate change impacts and economically disrupted and 1 otherwise. In this formulation, the value 1 is completely arbitrary: it could be replaced by any other positive number and perhaps discounted. This suggests that measures of how much decisions matter should be normalised

```
mMeas : (t : ℕ) → (n : ℕ) → X t → Double
mMeas t Z x     = 0.0
mMeas t (S m) x = let ps = bi (S t) m in
                  let v  = toDouble (val (bestExt ps :: ps) x) in
                  let v' = toDouble (val (worstExt ps :: ps) x) in
                  if v == 0 then 0 else (v − v') / v
```

Notice that, in states in which the control set is a singleton, any policy has to return the same control. In particular, the best extension and the worst extension of any policy sequence have to return the same control. Therefore, *mMeas* fulfils the avoidance condition of [BvH18] *per construction*. As a consequence, in S-states, the measure is always 0, independently of the time horizon:

```
∗Responsibility> : exec show (mMeas 0 4 (S, H, U))
"0"

∗Responsibility> : exec show (mMeas 0 6 (S, L, C))
"0"
```

Notice also that *mMeas* can be applied to estimate how much decisions matter at later steps of a decision process. For example, we can assess that, for our decision process and under a fixed time horizon, decisions in (D, H, U) at decision step 0 matter less than decisions in (D, H, U) at later steps:

```
∗Responsibility> : exec show (mMeas 0 7 (D, H, U))
"0.1730602684132721"

∗Responsibility> : exec show (mMeas 1 7 (D, H, U))
"0.5673067719100584"

∗Responsibility> : exec show (mMeas 3 7 (D, H, U))
"0.5673067719100584"
```

This is not surprising given that the best decision, in (D, H, U) and for a time horizon of 7 decision steps, is to start a green transition and that, as stipulated in the introduction and specified in Section 5 through

$$pSpec9 : p_{C|D,0} \leqslant p_{C|D}$$

the probability of entering states in which the world is committed to future severe impacts from climate change is higher in states in which a green transition has not already been started as compared to states in which a green transition has been started.

Wrap up. Following (S1), (S2) and (S3), we have introduced a measure of how much decisions under uncertainty matter that fulfils the requirements for responsibility measures put forward in the beginning of this section.

It accounts for all the knowledge which is encoded in the specification of a decision process, it is defined uniformly for any goal a (real or hypothetical) decision maker may pursue and it is *fair* in the sense that all decisions (decision makers) are measured in the same way.

Thus, we introduce *mMeas* as a first example of responsibility measure. In the next section, we generalise it by introducing a small DSL for the specification of goals of sequential decision processes under uncertainty and discuss alternative definitions.

7.2 A syntax for defining goals

Above we have defined the reward function in terms of a function

$$goal : \{t : \mathbb{N}\} \to (X\ t) \to \mathbb{B}$$

as

$$reward\ t\ x\ y\ x' = \mathbf{if}\ goal\ \{t = S\ t\}\ x'\ \mathbf{then}\ \ 0.0\ \mathbf{else}\ \ 1.0$$

In [BBC+21], we have defined a first simple DSL for the modular definition of such goals, allowing to put forward the goal of decision making in a transparent way. Using this syntax, the goal from above could be expressed as

$$goal = Avoid\ isCommitted\ \&\&\ Avoid\ isDisrupted$$

Here *Avoid* is a function that maps Boolean predicates to goals. It is one of the constructors of the abstract syntax

```
data Goal : Type where
  Exit   : Region → Goal
  Enter  : Region → Goal
  StayIn : Region → Goal
  Avoid  : Region → Goal
  (&&)   : Goal → Goal → Goal
  (||)   : Goal → Goal → Goal
  Not    : Goal → Goal
```

to specify goals for decision processes that are informed by notions of sustainable development or management [HKDM16b, Int18]: such goals are typically phrased in terms of a verb (*avoid*, *exit*, *enter*, *stay within*, etc.) and of a *region* (predicate, subset of states) that encode notions of planetary boundaries or operational safety[16].

In our formalisation, such regions are encoded by

```
Region : Type
Region = (t : ℕ) → Subset (X t)
```

where *Subset A* is an alias for $A \to \mathbb{B}$:

```
Subset : Type → Type
Subset A = A → 𝔹
```

Such regions might e.g. be assigned according to information obtained in commitment or tipping point computations with physical models (this will be further explored in TiPES D6.3, see also [MMCBB22b, MMCBB22a]).

The syntax allows moreover to combine goals with logical operators. Given the input of the framework's reward function (the current and successor states as well as the current control) and a goal, an evaluation function then assigns a semantic to the syntactic goal expression. The result of the evaluation is a Boolean value, indicating whether the goal is attained or not. The generic reward function simply does a case split on this Boolean value and returns a reward of 0 or 1.

[16] for example, in [HKDM16b], a partitioning of the state space into a *sunny* region and its *dark* complement is the starting point for the construction of a hierarchy of regions: shelters, glades, lakes, trenches and abysses, see figure 1 at page 7.

```
eval : Goal → (t : ℕ) → (x : X t) → Y t x → X (S t) → 𝔹
eval (Exit r)   t x y x' = let t' = S t in elem t x (r t)       ∧ ¬ (elem t' x' (r t'))
eval (Enter r)  t x y x' = let t' = S t in ¬ (elem t x (r t)) ∧ elem t' x' (r t')
eval (StayIn r) t x y x' = let t' = S t in                       elem t' x' (r t')
eval (Avoid r)  t x y x' = let t' = S t in                       ¬ (elem t' x' (r t'))
eval (g && g')  t x y x' = eval g t x y x' ∧ eval g' t x y x'
eval (g || g')  t x y x' = eval g t x y x' ∨ eval g' t x y x'
eval (Not g)    t x y x' = ¬ (eval g t x y x')
```

While the definition of *eval* is almost straightforward, domain experts do not need to be concerned with it. They just apply the constructors of *Goal* to specify the goal of decision making like in the definition of *goal* given above. The goal for which we measure how much decisions matter is then fully transparent and the rewards are a computed by a generic function based on *eval goal*:

```
goal : Goal
reward t x y x' = if eval goal t x y x'
  then 1.0
  else 0.0
```

In more realistic (as opposed to stylised, see section 6) GHG emissions decision processes, states are not necessarily either fully committed or fully uncommitted to severe impacts from climate change and decision makers are confronted with many degrees of commitment, possibly infinitely many.

A similar situation holds for other predicates on states, like being *vulnerable* (or *adapted*) to climate change or for measures of economic growth or welfare. This raises the question of how to specify the goals of decision making in decision processes in which predicates like *isCommitted* do not return Boolean values but, for example, values in $[0, 1]$. In this situation, a partitioning of the state space into regions is not immediately available and the specification of goals requires an extension both of the syntax *Goal* for encoding goals and of the interpretation function *eval* associated with this syntax.

Another possibility to generalise the simple DSL from above is to allow assigning weights to each atomic goal, likewise yielding rewards in the unit interval instead of just Boolean values:

```
UnitInterval : Type
UnitInterval = (d : Double ** So (0 ⊑ d ∧ d ⊑ 1))
castBD : 𝔹 → Double
castBD True  = 1.0
castBD False = 0.0

data Goal : Type where
  Exit   : UnitInterval → Region → Goal
  Enter  : UnitInterval → Region → Goal
  StayIn : UnitInterval → Region → Goal
  Avoid  : UnitInterval → Region → Goal
  (&&)   : Goal → Goal → Goal
  (||)   : Goal → Goal → Goal
  Not    : Goal → Goal

eval : Goal → (t : ℕ) → (x : X t) → Y t x → X (S t) → Double   -- better: UnitInterval
eval (Exit w r)   t x y x' = let t' = S t in
                               fst w * castBD (elem t x (r t) ∧ ¬ (elem t' x' (r t')))
eval (Enter w r)  t x y x' = let t' = S t in
                               fst w * castBD (¬ (elem t x (r t)) ∧ elem t' x' (r t'))
eval (StayIn w r) t x y x' = let t' = S t in fst w * castBD (elem t' x' (r t'))
eval (Avoid w r)  t x y x' = let t' = S t in fst w * castBD (¬ (elem t' x' (r t')))
eval (g && g')    t x y x' = min (eval g t x y x') (eval g' t x y x')
eval (g || g')    t x y x' = max (eval g t x y x') (eval g' t x y x')
eval (Not g)      t x y x' = 1 − eval g t x y x'
```

Not that the interpretation of the logical connectives is common for fuzzy logic [Zad75]. However, one might also want to explore the individual values for multiple goals at the same time without aggregating them.

$$
\begin{aligned}
&\mathit{w1} : \mathit{UnitInterval}\\
&\mathit{w1} = (0.4 \ast\ast\ \mathit{Oh})\\
&\mathit{w2} : \mathit{UnitInterval}\\
&\mathit{w2} = (0.3 \ast\ast\ \mathit{Oh})\\
\\
&\mathit{goal} : \mathit{Goal}\\
&\mathit{goal} = \mathit{Avoid}\ \mathit{w1}\ \mathit{isCommitted}\ \&\&\ \mathit{Avoid}\ \mathit{w2}\ \mathit{isDisrupted}\\
\\
&\mathit{reward}\ t\ x\ y\ x' =\\
&\quad \textbf{if}\ \mathit{eval}\ \mathit{goal}\ t\ x\ y\ x'\ \textbf{then}\ 1.0\ \textbf{else}\ 0.0
\end{aligned}
$$

Recall the function *trjR* of Section 3

$$\mathit{trjR} : \{t, n : \mathbb{N}\} \to \mathit{StateCtrlSeq}\ t\ n \to \mathit{List}\ \mathit{Val}$$

that computes a list with the rewards at each step. It is useful suitable for exploring the local fulfilment of goals. For such an exploration, it might be useful to preserve the information about the outcomes of multiple goals. We can implement this vectors of goals and vectors of doubles as type of values, instead of defining the reward function in terms of one single goal and resulting number:

$$
\begin{aligned}
&\mathit{dim} : \mathbb{N}\\
&\mathit{goals} : \mathit{Vect}\ \mathit{dim}\ \mathit{Goal}\\
\\
&Val = \mathit{Vect}\ \mathit{dim}\ \mathit{Double}\\
&\mathit{reward}\ t\ x\ y\ x' = \mathit{map}\ (\lambda g \Rightarrow \mathit{eval}\ g\ t\ x\ y\ x')\ \mathit{goals}
\end{aligned}
$$

Note that in order to compare such multi-goal rewards (for optimisation), they have to either be aggregated in some way or the comparison is performed according to a priority order on the components.

7.3 Some caveats

With *mMeas* defined as in section 7 and with *goal* : *Goal* specified as above, one can recover the results for the decision process of section 5. To wrap up, let us discuss a few aspects of the responsibility measures considered so far.

One important trait of these measures is that they are obtained by extending the decision process for which one wants to measure how much decisions matter to a fully specified finite horizon sequential decision problem. In comparison to approaches like those proposed in [Hal06], [CH04] and, more recently, [HH20], this approach has both advantages and disadvantages.

From the conceptual point of view, the major advantages are simplicity and straightforwardness: in contrast to models of causality like those put forward in the works mentioned above, finite horizon sequential decision problems are conceptually simple and well understood. Also, for finite horizon sequential decision problems, we can compute *provably* best and worst policies. This guarantees that the results obtained for a specific problem are a logical consequence of the assumptions made for that problem and not of programming errors or numerical errors. Because all the assumptions underlying a specific problem are put forward explicitly via specifications like

$$\mathit{goal} = \mathit{Avoid}\ \mathit{isCommitted}\ \&\&\ \mathit{Avoid}\ \mathit{isDisrupted},$$

the approach also guarantees high standards of transparency. Simplicity and straightforwardness are also the main drawbacks of our approach: we can only derive responsibility measures for decision processes that can be naturally extended to finite horizon sequential decision problems.

This is the case for the stylised GHG emissions decision process discussed throughout our work and, indeed, for many interesting problems in climate policy because, as pointed out in [Web08]:

> Climate policy decisions are necessarily sequential decisions over time under uncertainty, given the magnitude of uncertainty in both economic and scientific processes, the decades-to-centuries time scale of the phenomenon, and the ability to reduce uncertainty and revise decisions along the way.

But it is not immediately obvious how our approach could be applied to measure how much decisions matter in situations in which collective decisions emerge from a potentially large number of individual decisions, e.g., mediated through certain widely accepted mechanisms like majoritarian rules like in voting processes.

Another important aspect of the measures of responsibility proposed in this work is the comparison between verified best and what we called "conditional worst" decisions at the specific state at which we want to measure responsibility. Remember that, in the definition of *mMeas*, v and v' are $val\ (bestExt\ ps :: ps)\ x$ and $val\ (worstExt\ ps :: ps)\ x$, respectively. Here, $x : X\ t$ is a state at decision step t, ps is a verified optimal sequence of policies for taking n decisions starting from step $t + 1$ and $n + 1$ is the decision horizon.

Due to the definition of *bestExt*, generic backward induction and the correctness proof from section 4, *bestExt ps* :: *ps* is an optimal policy sequence and *bestExt ps* is an optimal policy (a function from states to controls) at decision step t. Similarly *worstExt ps* is a policy that guarantees

$$val\ (worstExt\ ps :: ps)\ x \sqsubseteq val\ (p :: ps)\ x$$

for all $x : X\ t$ and $p : Policy\ t$. In other words, we compare "best" decision (given by *bestExt ps*) and "worst" decision (given by *worstExt ps*) in x *conditional* to future decisions being best ones. This is crucial because the difference between best and worst decisions (and hence our estimates of how much decisions matter) at a given step and in a give state would in general be different if we assumed that future decision are not taken optimally.

In the context of our decision problem, for example, we would come up with a different measure of responsibility for "current" decisions if we assumed that future generations do not care about avoiding negative impacts from climate change or economic downturns or, equivalently, that they do care but do not act accordingly. If there are reasons to believe that this is the case, the verified optimal policy sequence *ps* in the definition of *mMeas* has to be replaced with one which is consistent with such a belief. For example, if we believe that the next generation will act more myopically (or more farsighted) than for a horizon of n decision steps, we have to compute *ps* accordingly.

Finally, we want to flag the role of the measure of uncertainty *meas* from section 2 in the definition of *val* and thus of v and v'. In all computations shown in this section, we have taken *meas* to be the *expected value measure* but other measures of uncertainty are conceivable and we refer interested readers to [Ion09, BJI17a] and [BB21].

8 Conclusion

We have developed domain-specific language elements for the study of dynamical systems and specification of SDP in the context of tipping point research. For better usability, we have implemented these language elements on top of a lightweight version of the original Botta et al. framework of [BJI17a]. We have defined generic measures of responsibility and a syntax to transparently express goals of decision making. We have illustrated the usage of the framework by specifying a conceptual stochastic green house gas emission problem. Furthermore, we have shown the correctness of the generic backward induction algorithm implemented in the framework in a more general setting than commonly considered in control theory. Considering that the aim of the current work is to improve accountability in the context of climate policy, this correctness result fills an important gap that had been overlooked in the previously existing verification result for the Botta et al. framework.

Nevertheless, many extensions to the work presented in this report are conceivable. Possible future work includes the following:

- Defining a syntax for systematic construction of transition functions from conditional probability functions, possibly following the approach of [Jac15, JZ20]

- Extending the language for value judgements (e.g. iterated notions of avoidance, reachability etc.), possibly following ideas for the topological classification of state spaces of [HKDM16b]
- Developing the algebraic theory of SDPs to improve modularity in the description of problems, extending work started in [KJ19].
- Incorporating notions for multi-objective / multi-stakeholder valuation for SDPs, following the risk-opportunity-analysis paradigm discussed in [MSV+20]
- Adding language elements for the description of problems with early warning signals (EWS) [BBCMM22c]

Acknowledgements

This report is part of TiPES deliverable D6.2. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 820970 (TiPES – Tipping Points in the Earth System, [TiP23]).

Many thanks go to Cezar Ionescu, Patrik Jansson, Tim Richter and Zheng Li who co-authored the papers [BBJR21] (P.J. and T.R.) and [BBC+21] (all) which were essential for the work presented in this report.

The work presented in this report heavily relies on free software, among others on Coq, Idris, Agda, GHC, git, vi, Emacs, LaTeX and on the FreeBSD and Debian GNU/Linux operating systems. It is our pleasure to thank all developers of these excellent products.

Conflicts of Interest

None.

References

[B+21] Nicola Botta et al. IdrisLibs. https://gitlab.pik-potsdam.de/botta/IdrisLibs, 2016–2021.

[B+22] Nicola Botta et al. IdrisLibs2. https://gitlab.pik-potsdam.de/botta/IdrisLibs2, 2019–2022.

[BB21] Nuria Brede and Nicola Botta. On the correctness of monadic backward induction. *Journal of Functional Programming*, 31:e26, 2021.

[BBC+21] Nicola Botta, Nuria Brede, Michel Crucifix, Cezar Ionescu, Patrik Jansson, Zheng Li, Marina Martínez-Montero, and Tim Richter. Responsibility under uncertainty: Which climate decisions matter most? *Submitted to Environmental Modeling & Assessment*, 2021.

[BBCMM20] Nicola Botta, Nuria Brede, Michel Crucifix, and Marina Martínez Montero. WP6 Course on functional languages and dependently typed languages: D6.1 (D27). https://doi.org/10.5281/zenodo.4543579, 2020.

[BBCMM21] Nicola Botta, Nuria Brede, Michel Crucifix, and Marina Martínez-Montero. A note on climate science and verified programming. https://doi.org/10.5281/zenodo.4543472, 2021.

[BBCMM22a] Nicola Botta, Nuria Brede, Michel Crucifix, and Marina Martínez-Montero. Decision theory and climate policy. https://doi.org/10.5281/zenodo.6817454, 2022.

[BBCMM22b] Nicola Botta, Nuria Brede, Michel Crucifix, and Marina Martínez-Montero. A note on climate science and climate policy. https://doi.org/10.5281/zenodo.6783575, 2022.

[BBCMM22c] Nuria Brede, Nicola Botta, Michel Crucifix, and Marina Martínez-Montero. Climate sensitivity, commitment and abrupt change: toward an ontology for climate tipping point research. https://doi.org/10.5281/zenodo.6820683, 2022.

[BBJR21] Nicola Botta, Nuria Brede, Patrik Jansson, and Tim Richter. Extensional equality preservation and verified generic programming. *Journal of Functional Programming*, 31:e24, 2021.

[BDLK18] Wolfram Barfuss, Jonathan F. Donges, Steven Lade, and Jürgen Kurths. When optimization for governing human environment tipping elements is neither sustainable nor safe. *Nature Communications*, 9:2354, 2018.

[Bel57] Richard Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[Ber95] P. Bertsekas, D. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, Mass., 1995.

[Bir14] R. Bird. *Thinking Functionally with Haskell*. Cambridge University Press, 2014.

[BJI17a] Nicola Botta, Patrik Jansson, and Cezar Ionescu. Contributions to a computational theory of policy advice and avoidability. *J. Funct. Program.*, 27:e23, 2017.

[BJI$^+$17b] Nicola Botta, Patrik Jansson, Cezar Ionescu, David R. Christiansen, and Edwin Brady. Sequential decision problems, dependent types and generic solutions. *Logical Methods in Computer Science*, 13(1), 2017.

[BJI18] N. Botta, P. Jansson, and C. Ionescu. The impact of uncertainty on optimal emission policies. *Earth System Dynamics*, 9(2):525–542, 2018.

[Bot20a] Nicola Botta. Bridging the gap between climate science and climate policy advice. https://doi.org/10.5281/zenodo.6783774, 2020.

[Bot20b] Nicola Botta. Climate science, program verification and policy advice. https://doi.org/10.5281/zenodo.4543621, 2020.

[Bot20c] Nicola Botta. The JFP2017 theory of verified policy advice: An overview. https://doi.org/10.5281/zenodo.4545679, 2020.

[Bot21] Nicola Botta. Responsibility under uncertainty: which climate decisions matter most? https://doi.org/10.5281/zenodo.6826502, 2021.

[Bot22] Nicola Botta. Decision problems in climate research, mathematical specification and dependent types. https://doi.org/10.5281/zenodo.6783894, 2022.

[Bra17] Edwin Brady. *Type-Driven Development in Idris*. Manning Publications Co., 2017.

[Bre20] Nuria Brede. Types for TiPES - applying type theory to climate impact research. https://doi.org/10.5281/zenodo.4554685, 2020.

[Bre21] Nuria Brede. Toward a DSL for Sequential Decision Problems with tipping point uncertainties. https://doi.org/10.5281/zenodo.6783894, 2021.

[Bre22] Nuria Brede. On the correctness of monadic backward induction. https://doi.org/10.5281/zenodo.6783894, 2022.

[BvH18] Matthew Braham and Martin van Hees. Voids or Fragmentation: Moral Responsibility For Collective Outcomes. *The Economic Journal*, 128(612):F95–F113, 01 2018.

[Car95] Rudolf Carnap. *Introduction to philosophy of science.* Dover, New York, 1995.

[CH04] Hana Chockler and Joseph Y. Halpern. Responsibility and blame: A structural-model approach. *J. Artif. Int. Res.*, 22(1):93–115, October 2004.

[Hal06] Joseph Y. Halpern. Causality, responsibility, and blame: A structural-model approach. In *Proceedings of the 3rd International Conference on the Quantitative Evaluation of Systems*, QEST '06, page 3–8, USA, 2006. IEEE Computer Society.

[Hal14] Joseph Y. Halpern. Cause, responsibility, and blame: A structural-model approach, 2014.

[HH20] Jobst Heitzig and Sarah Hiller. Degrees of individual and groupwise backward and forward responsibility in extensive-form games with ambiguity, and their application to social choice problems. *arXiv preprint arXiv:2007.07352*, 2020.

[HKDM16a] Jobst Heitzig, Tim Kittel, Jonathan F. Donges, and Nora Molkenthin. Topology of sustainable management of dynamical systems with desirable states: from defining planetary boundaries to safe operating spaces in the earth system. *Earth System Dynamics*, 7:21–50, 2016.

[HKDM16b] Jobst Heitzig, Tim Kittel, Jonathan F Donges, and Nora Molkenthin. Topology of sustainable management of dynamical systems with desirable states: from defining planetary boundaries to safe operating spaces in the earth system. *Earth System Dynamics*, 7(1):21–50, 2016.

[HL15] Ralf Hinze and Andres Löh. Guide to lhs2TEX, 2015.

[HWFD19] Koen G. Helwegen, Claudia E. Wieners, Jason E. Frank, and Henk A. Dijkstra. Complementing CO_2 emission reduction by solar radiation management might strongly enhance future welfare. *Earth System Dynamics*, 10(3):453–472, 2019.

[Int18] Intergovernmental Panel on Climate Change (IPCC). Ipcc, 2018: Summary for policymakers. https://www.ipcc.ch/site/assets/uploads/2018/10/SR15_SPM_version_stand_alone_LR.pdf, 2018.

[Ion09] Cezar Ionescu. *Vulnerability Modelling and Monadic Dynamical Systems.* PhD thesis, Freie Universität Berlin, 2009.

[Jac15] Bart Jacobs. New directions in categorical logic, for classical, probabilistic and quantum logic. *Logical Methods in Computer Science*, 11, 2015.

[JZ20] Bart Jacobs and Fabio Zanasi. The logical essentials of bayesian reasoning. *Foundations of Probabilistic Programming*, pages 295–331, 2020.

[KJ19] Robert Krook and Patrik Jansson. An algebra of sequential decision problems. Technical report, Technical Report. Computer Science and Engineering, Chalmers University of . . . , 2019.

[Mit97] Thomas M. Mitchell. *Machine Learning.* McGraw-Hill, Inc., USA, 1 edition, 1997.

[MMB20] Marina Martínez Montero and Nuria Brede. Can Solar Radiation Management help to avoid greenland's tipping point? https://doi.org/10.5281/zenodo.4554708, 2020.

[MMCBB22a] Marina Martínez Montero, Michel Crucifix, Nicola Botta, and Nuria Brede. Commitment as lost opportunities. Technical report, Copernicus Meetings, 2022.

[MMCBB22b] Marina Martínez Montero, Michel Crucifix, Nicola Botta, and Nuria Brede. Commitment as lost options. Upcoming., 2022.

[MSV$^+$20] Jean-Francois Mercure, Simon Sharpe, Jorge Vinuales, Matthew Ives, Michael Grubb, Hector Pollitt, Florian Knobloch, and Femke Nijsse. Risk-opportunity analysis for transformative policy design and appraisal. 2020.

[Nor18] William Nordhaus. Evolution of modeling of the economics of global warming: changes in the DICE model, 1992–2017. *Climatic Change*, 149(4):623–640, 2018.

[ODC$^+$20] Ilona M. Otto, Jonathan F. Donges, Roger Cremades, Avit Bhowmik, Richard J. Hewitt, Wolfgang Lucht, Johan Rockström, Franziska Allerberger, Mark McCaffrey, Sylvanus S. P. Doe, Alex Lenferna, Nerea Morán, Detlef P. van Vuuren, and Hans Joachim Schellnhuber. Social tipping dynamics for stabilizing earth's climate by 2050. *Proceedings of the National Academy of Sciences*, 117(5):2354–2365, 2020.

[Pin17] Robert S. Pindyck. The Use and Misuse of Models for Climate Policy. Review of Environmental Economics and Policy, 2017.

[Put14] Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[RSN$^+$09] J. Rockström, W. Steffen, K. Noone, A. Persson, F. S. Chapin, E. Lambin, T. M. Lenton, M. Scheffer, C. Folke, H. Schellnhuber, B. Nykvist, C. A. De Wit, T. Hughes, S. van der Leeuw, H. Rodhe, S. Sörlin, P. K. Snyder, R. Costanza, U. Svedin, M. Falkenmark, L. Karlberg, R. W. Corell, V. J. Fabry, J. Hansen, B. Walker, D. Liverman, K. Richardson, P. Crutzen, and J. Foley. Planetary boundaries: Exploring the safe operating space for humanity. *Ecol. Soc.*, 14(2):32, 2009.

[She19] Theodore G. Shepherd. Storyline approach to the construction of regional climate change information. *Proc. R. Soc.*, 475(2225), 2019.

[SMV$^+$21] S. Sharpe, J-F. Mercure, J. Vinuales, M. Ives, M. Grubb, H. Pollitt, F. Knobloch, and F.J.M.M. Nijsse. Deciding how to decide: Risk-opportunity analysis as a generalisation of cost-benefit analysis. Technical report, UCL Institute for Innovation and Public Purpose, Working Paper Series (IIPP WP 2021/03), 2021.

[The10] The Idris Community. The Idris Tutorial (revision 0417c53f), 2017-2010.

[TiP23] TiPES H2020 Project Website. https://www.tipes.dk/, 2019–2023.

[Wad92] Philip Wadler. Monads for functional programming. In *Program Design Calculi, Proceedings of the NATO Advanced Study Institute on Program Design Calculi, Marktoberdorf, Germany, July 28 - August 9, 1992*, pages 233–264, 1992.

[Web08] Mort D. Webster. Incorporating Path Dependency into Decision-Analytic Methods: An Application to Global Climate-Change Policy. *Decision Analysis*, 5(2):60–75, 2008.

[Zad75] Lotfi A Zadeh. Fuzzy logic and approximate reasoning. *Synthese*, 30(3):407–428, 1975.

A note on climate science and climate policy

The TiPES-WP6 team

2021-06-29

Doing the right things

In its fourth Assessment Report (Intergovernmental Panel on Climate Change (IPCC) 2007), the Intergovernmental Panel on Climate Change (IPCC) has pointed out that responding to climate change involves "*an iterative risk management process that includes both mitigation and adaptation, taking into account actual and avoided climate change damages, co-benefits, sustainability, equity and attitudes to risk.*"

Fifteen years later, we have to recognize that humanity is far from having implemented such an "iterative risk management process", that our scientific understanding of the notions involved in this process is less than satisfactory and that solutions towards keeping the earth climate within safe boundaries are difficult to agree upon and implement in practice.

This is not very surprising if one considers that different decision makers, say, countries or coalitions between countries, are due to experience different (negative and positive) impacts from climate change, are in very different cultural, economic and technological situations and, perhaps more importantly, are in competition (if not in war) with each other.

What is perhaps more surprising is that, even within the scientific community, there is little agreement on how to turn the scientific knowledge distilled in the IPCC Assessment Reports into advice to policy makers that is *pragmatic*, *transparent* and, above all, *accountable*.

Doing the right things rightly

In spite of a strong focus on quantitative analysis and prediction, climate science has been so far embarrassingly incapable of providing advice on matters of climate policy that is *accountable*: decision makers do not precisely know what kind of outcomes and guarantees they can expect from implementing the advice received.

This, too, is not very surprising. Applications of the physical sciences, e.g., to engineering or to public health, heavily rely on *empirical* methods. Where predictions are necessarily uncertain – e.g., because of the lack of well established theories or because of imperfect information – the physicist (chemist, biologist, etc.) can often turn to experiments, either in a laboratory or on the field.

The evidences obtained in such tests and experiments are recorded in formal protocols, analysed and perhaps confirmed (or confuted) by other experiments and finally applied to *pragmatic* decision making.

While formal methods cannot fully replace empirical verification, they can provide very high levels of *transparency* and contribute towards making political decisions more understandable, more transparent and more accountable (Botta et al. 2020).

Differences that matter

However, applying formal methods requires both advisors *and* decision makers to have achieved a shared understanding of the impacts of uncertainties on decision making (Botta, Jansson, and Ionescu 2018), of the differences between decision making under uncertainty and decision making in a deterministic environment and to carefully distinguish between closely related but crucially different notions. In this section, we discuss some of the differences that matter.

Acting vs. planning.

You have had breakfast and are on your way to your office. You drive the car out of the garage, fire up Google Maps on your mobile phone, enter your position and select your office as your goal. You are suggested a route, start driving and follow the suggestions of the routing algorithm. On your way to the office you get re-routed a couple of times, perhaps because of an accident on the original route or because you have made a detour to pick-up a colleague who has called you while you were driving.

In following or rejecting the recommendations of the routing algorithm, you are taking decisions, one after the other. Some of the these decisions entail judgments about uncertain events. Perhaps if you pick-up the colleague you might be caught in a traffic jam and miss an important meeting.

At each decision step, you are concerned with making a *best* decision, one that will get you to the office in the shortest time. Or, perhaps, one that is safest or one that avoids driving through a district you hate.

No matter what your aims are, at each decision step you want to take a decision that best matches your aims. Google Maps is your friend and you have learned how to judge its advice. You start with a route than you trust being the best possible given the information available at the time you drive out of the garage. Perhaps you revise your original plan on your way to the office. For instance, if Google Maps suggests you an alternative route.

In driving and taking decisions on the way to your office, you are *acting.* In doing so, you are exploiting the results of another activity: *planning.*

Planning and acting are closely related but essentially different activities. While driving to the office, driving decisions follow a plan. But the plan evolves in time, following the decisions.

While planning and acting may take place simultaneously, they are *logically distinct* activities. Sometimes, like in the example of driving to the office, planning and acting are concerns of two different agencies: Google Maps is responsible for planning, you – the driver – for acting.

Often, the same agent is involved in planning and acting, typically at different times.

Another example: tomorrow we want to bike to the countryside. We plan a long tour but the weather forecast is uncertain. In the morning it should be sunny but in the afternoon there is a significant chance of thunderstorms. Thunderstorms will come from west and they might align. In that case, we might get heavy hail.

We pack our rain clothes into the bicycle bags but we plan to break the tour if the weather gets really bad by 3pm. We will take a slightly longer route that will allow us to easily reach a train station if we decide to break the tour before 3pm. In this case, we will come back by train. If tomorrow morning the weather forecast worsens, we will make a short tour instead and be back for lunch.

We have made a plan for two decision steps, one tomorrow morning and one tomorrow afternoon. For each step, we have defined a decision rule: for each possible *state* (in step one, same/worse forecast; in step two, weather stable/really bad) we have planned a corresponding *action*.

Planning under uncertainty

Thus, we have defined two functions, one for each step. Tomorrow morning we will check the weather forecast, apply the first function and decide whether we go for a long or for a short ride.

It is important to realize that, under uncertainty, planning essentially means defining decisions *functions*, one for each decision step. In control theory, these functions are called *policies*. In game theory, they are often called *strategies* or sometimes *contingency plans* (Puterman 2014).

Thus, when we speak of *optimal* plans for a specific decision problem (no matter what optimal means for that specific problem), we speak of optimal *policy sequences*.

This is in contrast to planning for *deterministic* decision problem that is, for decision problems without uncertainty. In this case plans (and, therefore, optimal plans) can be conceived as sequences of actions.

This is because, in absence of uncertainties, a decision at a given step uniquely defines the conditions under which the next decision step takes place.

It goes without saying that, in most realistic situations, planning takes place under uncertainty. Ignoring uncertainties can lead to inefficiencies and fragile planning, as sometimes observed in planned economies.

In *practical* climate decision problems, decisions are taken sequentially and uncertainties are typically unavoidable (Webster 2000), (Webster 2008). They are a consequence of imperfect scientific knowledge but also, and more importantly, of political instability, inertia of legislations and of the intrinsic uncertainty of technological innovation.

Even if we assumed a perfect scientific knowledge of the processes that determine the impacts of GHG emissions on the climate, planning for GHG emission problems would still have to account for these uncertainties.

From this angle, speaking of “emission paths” (in contrast to emission policies or, perhaps more explicitly, of emission decision functions) suggests a fundamental misunderstanding of the problem at stake: no “optimal” emission path can be a meaningful answer to the problem of planning “good” decisions in, e.g. solar radiation management problems (Moreno-Cruz and Keith 2012), (Helwegen et al. 2019), (Nordhaus 2019).

As planning under uncertainty means defining *policies* that is, decisions functions, for each decision step, what are the domains and the codomains of such functions?

For a given decision step, the *domain* of a policy consists of the set of the observations that can be done at that step and that are relevant to decision making.

For tomorrow's morning decision step, we have contemplated only two possible observations: that the weather forecast has worsened or that the weather forecast is unchanged. Perhaps we should also consider the possibility that the weather forecast improves and, in that case, leave our rain clothes at home. We might want to make even more realistic plans and consider the possibility that tomorrow morning we feel very tired or lazy and decide to stay home no matter how the weather will be. No matter how the possible observations looks like, in decision theory, they are called the set of possible *states*.

The *codomain* of a decision function – a *policy* – is typically different in different states. In a given state, it consists of all the actions (options) that can be done in that state.

In our plan for tomorrow, the options are to go for the short tour or to go for the long one in both states. The policy that will guide our decisions is to go for the long tour if the weather forecast is unchanged and for the short ride if it has worsened.

In control theory, the set of actions (options) considered in a given state is called the *controls* set for that state.

Acting under uncertainty: optimality and regret.

We have seen that planning under uncertainty means finding sequences of policies or, in other words, sequences of decision functions.

Sometimes, we can estimate the (uncertain) consequences of acting according to a fixed sequence of decision functions. For instance, we can compute the *possible trajectories* associated with taking decision according to the policy sequence and perhaps even their probabilities.

If we are also able to attach values to possible trajectories, we can often compute so-called *optimal policies*.

The measure of uncertainty accounts for how decision makers aggregate the (uncertain) values associated with the possible trajectories. For example, a risk-neutral decision maker might measure stochastic-uncertainty according to the *expected-value measure*. In the same situation, a risk-averse decision maker might adopt a measure that minimizes the probability of worst outcomes.

What can a decision maker expect from actually taking decisions according to an optimal sequences of decision functions? Can optimality avoid **regret**?

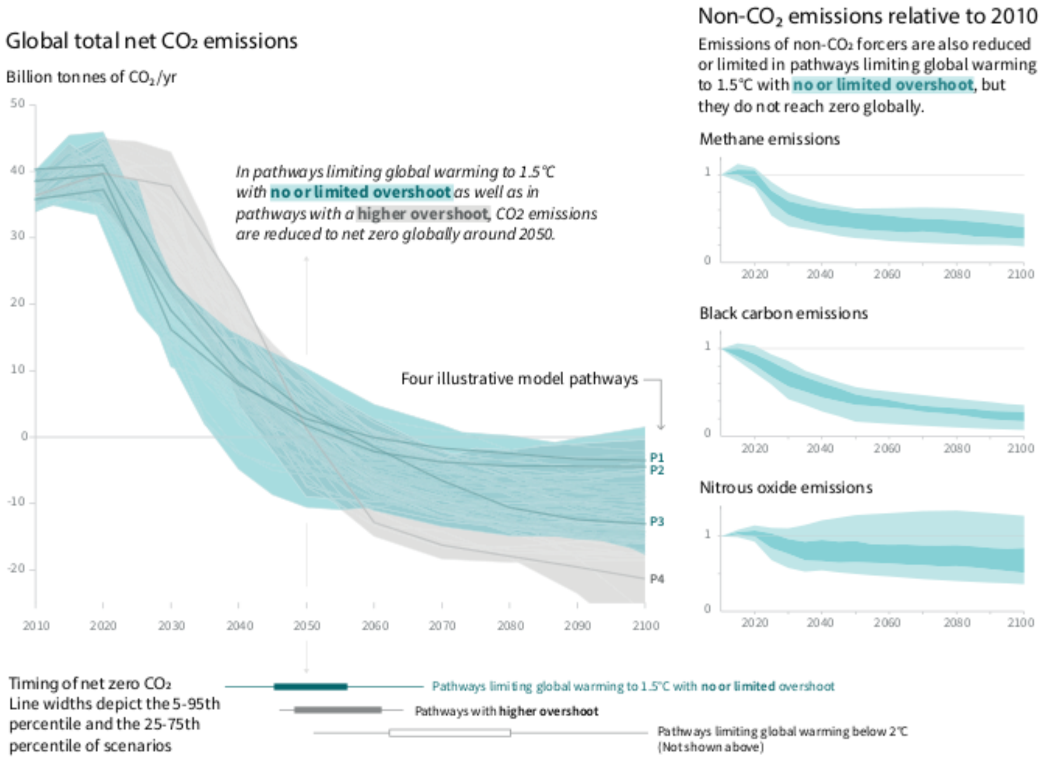
Unfortunately, this is not the case. Even if we follow provably optimal decision rules, we can always have bad luck and take a decision that in hindsight we might regret: avoiding smoking and regularly go biking does not guarantee one not to die from lung cancer. Still, it's a better policy than chain smoking and sitting the whole day in front of a computer.

This is another important difference between decision making under uncertainty and decision making in a deterministic environment: in the deterministic case optimal decisions do indeed guarantee regret-free decision making.

What to do and how to do it.

Another difference that must be kept in mind when considering the problem of applying climate science to policy making is that between what to do and how to do it.

Consider, for example, the guideline on global GHG emissions at page 19 of the summary for policy makers of the IPCC special report on global warming of 1.5 °C (Intergovernmental Panel on Climate Change (IPCC) 2018):



The blue corridor entails emission paths that, according to the knowledge available at the point in time in which the summary was prepared, limit global warming to 1.5 °C with no or limited overshooting.

The summary and, specifically, the corridor provides crucial information to decision makers. However it does not attempt at answering the question of how to actually implement an emission path that is consistent with the “safe” emission corridor. Answering this question has very different dimensions that can hardly be covered within climate science.

Along one such dimensions we have the problem of finding sequences of policies (or decision rules, see section Planning under uncertainty) that support pragmatic decisions (e.g., on GHG abatement targets at a given point in time and in a given state) that are likely to yield global GHG emissions within the “safe” corridor. The focus here is on *likely*: global decision are necessarily uncertain (Rougier and Crucifix 2018) and every attempt at finding realistic policy sequences has to account for such uncertainties. Tackling the problem of finding policy sequences under uncertainty requires contributions from, among others, control theory, expert elicitation, computer science and of course climate science (Webster 2000), (Botta, Jansson, and Ionescu 2017), (Webster 2008), (Helwegen et al. 2019), (Heitzig et al. 2016), (Botta, Jansson, and Ionescu 2018).

Another obvious dimension of the problem of applying global guidelines to policy making entails the question of how to actually get decision makers (countries) that are likely to experience different (negative and positive) impacts from climate change and that are in competition with each other to actually coordinate and cooperate to achieve global goals. The question is at the border between moral philosophy (Hardin 1968), (Ockenfels, Werner, and Edenhofer 2020), game theory (Heitzig 2012) and economics. Recently formal methods have been proposed as a means of improving the accountability of mechanism (rules) that are designed to fulfill well defined specifications (Caminati et al. 2015), (Rowat, Kerber, and Lange-Bever 2016).

Finally, a crucial dimension of the *how to do it* problem is technological: is it meaningful to complement unavoidable GHG emissions reductions with solar radiation management measures? Will nuclear fusion and GHG sequestration arrive in time to mitigate the impacts of fossil fuel economies?

The bottom line

At the interface between climate science and climate policy there is plenty of opportunities for confusion and misunderstandings. We have flagged differences that matter and that is worth keeping in mind when discussing how to turn scientific knowledge into advice to policy makers.

References

Botta, Nicola, Nuria Brede, Michel Crucifix, Cezar Ionescu, Patrik Jansson, and Marina Martínez. 2020. "Climate Science and Verified Programming." TiPES WP6 internal note.

Botta, Nicola, Patrik Jansson, and Cezar Ionescu. 2017. "Contributions to a Computational Theory of Policy Advice and Avoidability." *J. Funct. Program.*, nos. 27, e23.

———. 2018. "The Impact of Uncertainty on Optimal Emission Policies." *Earth System Dynamics* 9 (2): 525–42. https://doi.org/10.5194/esd-9-525-2018.

Caminati, Marco B., Manfred Kerber, Christoph Lange-Bever, and Colin Rowat. 2015. "Sound Auction Specification and Implementation." In *EC '15 Proceedings of the 16th ACM Conference on Economics and Computation*, edited by Tim Roughgarden, Michal Feldman, and Michael Schwarz, 547–64. Association for Computing Machinery. https://doi.org/10.1145/2764468.2764511.

Hardin, Garrett. 1968. "The Tragedy of the Commons." *Science* 162 (3859): 1243–8. http://www.sciencemag.org/cgi/content/full/162/3859/1243.

Heitzig, Jobst. 2012. "Bottom-Up Strategic Linking of Carbon Markets: Which Climate Coalitions Would Farsighted Players Form?" *SSRN Environmental Economics eJournal.* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2119219.

Heitzig, Jobst, Tim Kittel, Jonathan F. Donges, and Nora Molkenthin. 2016. "Topology of Sustainable Management of Dynamical Systems with Desirable States: From Defining Planetary Boundaries to Safe Operating Spaces in the Earth System." *Earth System Dynamics* 7: 21–50.

Helwegen, Koen G., Claudia E. Wieners, Jason E. Frank, and Henk A. Dijkstra. 2019. "Complementing CO_2 Emission Reduction by Solar Radiation Management Might Strongly Enhance Future Welfare." *Earth System Dynamics* 10 (3): 453–72. https://doi.org/10.5194/esd-10-453-2019.

Intergovernmental Panel on Climate Change (IPCC). 2007. "Climate Change 2007: Synthesis Report. Contributions of Working Groups I, II and III to the Fourth Assessment Report of the

Intergovernmental Panel on Climate Change."

———. 2018. "IPCC, 2018: Summary for Policymakers." Edited by V. Masson-Delmotte, P. Zhai, H. O. Pörtner, D. Roberts, J. Skea, P. R. Shukla, A. Pirani, et al. https://www.ipcc.ch/site/assets/uploads/2018/10/SR15_SPM_version_stand_alone_LR.pdf; Intergovernmental Panel on Climate Change (IPCC).

Moreno-Cruz, Juan, and David Keith. 2012. "Climate Policy Under Uncertainty: A Case for Geoengineering." *Climatic Change*. https://doi.org/10.1007/s10584-012-0487-4.

Nordhaus, William. 2019. "Economics of the Disintegration of the Greenland Ice Sheet." *Proceedings of the National Academy of Sciences* 116 (25): 12261–9. https://doi.org/10.1073/pnas.1814990116.

Ockenfels, Axel, Peter Werner, and Ottmar Edenhofer. 2020. "Pricing Externalities and Moral Behaviour." *Nature Sustainability*, 361–80. https://doi.org/10.1038/s41893-020-0554-1.

Puterman, Martin L. 2014. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Rougier, Jonathan, and Michel Crucifix. 2018. "Uncertainly in Climate Science and Climate Policy." *arXiv: Physics and Society*, 361–80.

Rowat, Colin, Manfred Kerber, and Christoph Lange-Bever. 2016. "An Introduction to Mechanized Reasoning." *Journal of Mathematical Economics* 66 (October): 26–39. https://doi.org/10.1016/j.jmateco.2016.06.005.

Webster, Mort D. 2000. "The Curious Role of "Learning" in Climate Policy: Should We Wait for More Data?" MIT Joint Program on the Science; Policy of Global Change, Report No. 67.

———. 2008. "Incorporating Path Dependency into Decision-Analytic Methods: An Application to Global Climate-Change Policy." *Decision Analysis* 5 (2): 60–75.

A note on climate science and verified programming

The TiPES-WP6 team with Cezar Ionescu and Patrik Jansson

Parts of this note have been taken almost verbatim from (Ionescu et al., 2018) and (Botta et al., 2017).

Verified programming: what is that?

The approach of work package 6 "Understand and communicate the impacts of Tipping Point uncertainties on accountable policies" (WP6) of the H2020 EU TiPES "Tipping Points in the Earth System" project is based on three pillars: *climate science*, *decision theory* and ***verified programming***.

Most climate scientists will agree that understanding and communicating the impacts of tipping point uncertainties on climate policies requires contributions from climate science. And many will agree that understanding the impacts of uncertainties on climate policies (no matter whether these are about the presence or the magnitude of abrupt transitions or about the value of model parameters at which structural changes in relevant features of the model take place) requires some understanding of the decision processes for which such policies are envisaged and, hence, of decision theory[1].

But what has *verified programming* to do with all this? What has verified programming to do with climate science, and, most importantly, what *is* verified programming and ***who needs it***?

Who needs verified programming?

In a nutshell, verified programming is a methodology for writing programs that can be checked to be correct by a *type checker.*

A type checker is itself a program (perhaps written in a language that is not that of the programs it checks) and a program is correct if it fulfills a *specification.* For example, a specification for a program R that is meant to compute square roots of positive real numbers might look like

$$\forall x \in \mathbb{R}, \quad 0 \leq x \Rightarrow R(x) * R(x) = x \tag{1}$$

Some machinery is needed in order to turn (1) into a *formal* statement that a type checker can actually process. Also, the notion that a square root program R is correct if it can be shown to

[1] Decision theory is a rather broad notion. Accountable climate policy advice necessarily requires contributions from different disciplines including, among others, control theory, expert elicitation and game theory. In TiPES WP6 we will follow (Webster, 2000), (Webster, 2008) and focus of control theory and sequential decision problems under non-deterministic and stochastic uncertainty, see (Botta et al., 2020).

fulfill (1) by a type checker is not without problems: what if the type checker itself is incorrect? And what does this mean? These are interesting and relevant question but we do not need to be concerned with them here.

For the purpose of this discussion, we only need to realize that (1) assigns a *meaning* to R. It is a precise and yet concise, description of what R is required to do. It is certainly more concise and more understandable than R itself can possibly be, especially when R is written in an imperative programming language[2].

If the type checker verifies R to fulfill (1), we can be sure that, as long as x is non-negative and R terminates on x, $R(x)$ is a square root of x.

This is quite something but if we are paying a lot of money for somebody to implement R, we might want to get a little bit more for our money. First, we might want R to always terminate, or, at least, to terminate for values of x in a suitable range. Second, we might want to make sure that R always delivers a positive root. Nothing in our specification so far prevents R to deliver -1 for 1 and 2 for 4!

If we demand more from R, we will typically have to pay more as the implementor will face a more difficult task. The *strength* of a specification is a crucial trait of the contract between the client of a program, say P, and the developer of P. The latter is always free to deliver a program that meets stronger requirements S' than those agreed on in the specification S

$$S'(P) \Rightarrow S(P) \tag{2}$$

Symmetrically, the client is free to accept programs that deliver less than agreed on:

$$S''(P) \Leftarrow S(P) \tag{3}$$

But delivering less than what has been agreed on in the specification (or requiring more) is a potential source of conflicts and perhaps court disputes.

Another source of potential misunderstanding are *impossible* specifications. Often, the client simply demands too much. In the case of the square root function, for instance, demanding $R(x) * R(x)$ to be exactly equal to x is too much if R has to terminate and can thus only compute *approximations* of irrational roots. Impossible specifications are another potential source of misunderstandings and should be avoided. Program that fulfill impossible specifications are often called *miracles*, see (Morgan, 1990).

[2] With a slight oversimplification, there are two distinct families of programming languages: imperative and functional. Examples of imperative languages are FORTRAN, C, C++, Java. Examples of functional languages are Haskell, Agda, Coq (The Coq Development Team, 2020), Idris. Imperative programming is a method of specifying what a computing machine shall do in terms of *instructions* and *execution procedures*. In functional programming, one specifies what a computing machine shall do in terms of *functions* and their *application* and *composition*, with an emphasis on inductive definitions and algebraic structure. In functional languages, the expression `a = b` has the same meaning as in mathematics. In imperative programming this is not the case and instructions like `a = 1; a = 3` are valid in spite of the fact that 1 is not equal to 3.

The discussion above should have made clear that verified programming is also (perhaps mainly) about being precise about the computations that a program shall perform. As a first step, this is done by putting forward mathematical specifications. In turn, these assign precise meanings to programs.

Indeed, one of the first papers on verified programming was Floyd's 1967 "Assigning Meaning to Programs" paper (Floyd, 1967). It is one of the seminal papers in computer science that still inform modern program verification.

Today, all programs that somehow matter – program that control medical equipment, financial transactions, weapons, access to critical data, power plants, air control systems, etc. – rely on some form of formal verification.

But who does actually need verified programming in science? What does it have to do with scientific computing? Do numerical analysts need to verify their programs? What about climate scientists?

It is probably fair to say that, as long as scientists are operating in a purely academic environment, they do not need to care about program verification: no university teacher is likely to loose her job because of a programming error.

Still, there are prominent examples of scientific claims that have been founded on programming errors (no citations here!). And in absence of clear, unambiguous specifications, even careful physical experiments and testing can easily lead to severe, regrettable consequences (Wikipedia, 2020b), (Lions and others, 1996), (Wikipedia, 2020a).

So do we all need verified programming? The answer is yes, but at different dosages.

As long as we are working on problems that are very well understood, program verification does probably not need to be our major concern. Precise specifications could still save us a lot of tedious work and time but, as long as we are implementing a new discretization for the Navier-Stokes equation, perhaps one that accounts for some insights from asymptotic analysis, we can rely on a whole body of knowledge and theoretical understanding of the problem at stake. In these cases, we indeed rely on very precise, albeit in most cases implicit, specifications. The same holds when our program is meant to deal with stiff ordinary differential equations or when we are implementing multi-grid methods for solving elliptic partial differential equations.

Things start to become different when we move from numerical methods for, e.g., the Euler equations to numerical methods for weather prediction or, even worse, global circulation models (GCM).

Here, the air starts to become thinner and our safety network less reliable. We can try to compensate for the lack of general results with careful testing. But tests can only show the presence of errors, not their absence: in front of unexpected results and without verified programs, we cannot know whether we are confronted with model deficiencies or with errors in the implementations of the models. In this situation, model validation becomes impossible.

Things get worse when we move from GCMs to intermediate complexity models and at the latest when we get to integrated assessment models or, even worse, non-deterministic or stochastic agent-based models or models for decision making, program verification becomes mandatory.

But is it not enough to test our models? Cannot we simply test our square root program R on a sample of inputs that is representative of the values for which we want to compute square roots? We answer this question in the next section.

Testing vs. proving

There are basically two methodologies for assessing that a program behaves according to a specification: testing and proving (Ionescu and Jansson, 2013).

In testing, a program is required to pass a finite number of tests in order to be positively verified. In proving, the program has to be shown to fulfill a formal specification.

In engineering and industrial applications, testing is well supported (Claessen and Hughes, 2000). It has a strong historical record of successes interspersed with a few dramatic failures (Wikipedia, 2020b), (Lions and others, 1996), (Wikipedia, 2020a).

Testing and proving are complementary methods. Testing can show the presence of errors, proving can show their absence. When can we test, when do we have to prove?

Discussing these questions goes beyond the scope of this note, but notice another crucial difference between testing and proving: testing a program P requires running P. By contrast, proving that P fulfills a specification does not require running P.

Thus, when available, proving is the method of choice when running a program takes a lot of time or is very expensive or dangerous. The other way round, when running a program is cheap and safe, testing is a viable choice.

It is also worth noticing that there are cases in which testing and proving are equivalent and thus, testing can indeed ascertain the absence of errors. Can you see when this is the case?

No matter whether we are trying to assess the correctness of a program by tests or formal proofs, we always need a specification. For our square root program R, for instance, we need something like (1). In absence of specifications, we do not know how to test R and also we do not know how to prove that R is correct.

This note is about verified programming and we are not insisting on testing here. However, given the importance of modelling in climate science, let us flag the role of design-by-contract as a method for developing and testing models.

In a nutshell, design-by-contract is a method for encoding program specifications in run-time tests. The methodology has been popularized in the late eighties, mainly through the work of Bertrand Meyer (Meyer, 1986), (Meyer, 1997), (Meyer, 1992), see also (Meyer, 2020). It allows programmers to specify and document programming tasks and to detect failures to comply with the specification at run time.

If you write programs in, among others, D, Eiffel, Fortress, Scala or Clojure, you can rely on native support for design-by-contract patterns. If you code in C, C++, Java or Python, check language-specific libraries for contract, e.g., (Caminiti, 2020) for C++. For other programming languages, see examples of design-by-contract patterns implemented via assertions on (RosettaCode, 2020).

Design-by-contract cannot guarantee that programs are correct. But it can signal the presence of errors and this has lead to significantly faster and more understandable software development.

A further step toward building programs from verified components has been achieved through Quickcheck (Claessen and Hughes, 2000). Quickcheck is a *combinator* library. It has been designed to assist program testing and is available in most programming languages, see (Wikipedia, 2020c).

Proving: verified programming

How do we actually verify that a program fulfills a specification? This is typically done within a *specification language* and using computer-assisted formal methods.

Until about two decades ago, programs were written and specified in different languages: programming languages and specification languages like Z (Bowen, 1996), VDM (Jones, 1990), B (Abrial, 1996) or Maude (Clavel et al., 2007).

Today, we can rely on a unified framework for program specification and program implementation, one that is mature (several decades old), with solid implementations (NuPRL (Allen et al., 2006), Coq (The Coq Development Team, 2020), Agda (Norell, 2007), Idris (Brady, 2017), Lean (de Moura et al., 2015)), and impeccable mathematical credentials: *Dependent Type Theory*.

In short, Dependent Type Theory (in the following just "Type Theory") is a pure functional programming language with a static type system. It is similar to Haskell (Kees Doets and Eijck, 2004), (Bird, 2014), and stands in roughly the same relation to it as predicate logic to propositional logic (Moschovakis, 2018). Type Theory was developed by the Swedish mathematician and philosopher Per Martin-Löf (Martin-Löf, 1984), who intended it to have the same foundational role for intuitionistic mathematics that set theory expressed in predicate logic had for classical mathematics.

This is not the place for a presentation of Type Theory, for a particularly accessible one, see (Altenkirch, 2017). What we want to do here is to provide an intuition for why Type Theory provides an environment for both program specification and program implementation and for how this environment is used in program verification.

We start by recalling that set theory derives its foundational role in classical mathematics from its ability to represent properties in several different (equivalent) ways, within a first-order language. For example, given a property P over a set A, expressed as a formula in the first-order language of sets, we can view it as a

- set $P = \{a \mid P\ a\}$, $a \in P$ iff a has the property P
- Boolean-valued function: $P : A \to Bool$, $P\ a = True$ iff a has the property P
- set-valued function: $P : A \to \{\{\}, \{*\}\}$, $P\ a = \{*\}$ iff a has the property P

All these allow us to talk about the property **within** the theory: it becomes an element of the universe of discourse. If we take types in programming languages to be the analogues of sets in set theory, we can see that the available means for their construction are more restricted. In common with other functional programming languages, Type Theory allows the construction of inductive types. For example

$$\frac{}{Z : Nat} \qquad \frac{n : Nat}{S\ n : Nat}$$

and

```
data Nat : Type where
  Z : Nat
  S : Nat -> Nat
```

are two equivalent ways of expressing the familiar rules for the inductive construction of the natural numbers: zero (Z, `Z`) is a natural number; if `n` is a natural number then the successor of `n` ($S\ n$, `S n`) is a natural number, etc.

The first definition is written using *inference rules* - this is how the rules of logical proof systems are commonly presented (Plato, 2018); the other one is written in the style of Haskell, Agda, or Idris. In most programming languages, we can represent properties as Boolean-valued predicates. For example in Haskell:

```
isEven : Nat       ->  Bool
isEven Z           =   True
isEven (S Z)       =   False
isEven (S (S m))   =   isEven m
```

In most cases, however, we cannot represent the associated set as a datatype or as a type-valued function. Therefore, if a function requires its argument to be even, then the best we can do is to guard the call of the function with a run-time test. This leads to expressing requirements or specifications as tests, as in *test-driven development* methods or, as discussed above, design-by-contract.

In contrast, in Type Theory, we have the *additional* possibility of representing a property by a type-valued function (a type *family*), which corresponds to the set-valued version in set theory. For example

```
data Even : Nat -> Type where
  MkEven : (k : Nat) -> Even (2 * k)
```

is a way of expressing the type-valued function version of `isEven`. For every natural number `n`, `Even n` is a type. If `n` is not even, then the type will be empty. Otherwise, the type will have one element, namely `MkEven (n / 2)`.

If a function requires its argument to be even, we can now formulate this requirement at the level of its type, for instance

```
f : (n : Nat) -> Even n -> X
```

In order to call `f` with an argument `n`, we have to supply another argument of type `Even n`. We can only do that if `n` is `Even`, since otherwise `Even n` would be empty. This additional argument must be reducible to the form `MkEven k`, where `k = n / 2`, and this can be checked at *compile time* (or, rather, at “type-checking time”). This ensures that `f` will never give rise to a run-time error, a much stronger guarantee than we can enforce by means of tests.

The ability to define inductive data types and type families lends Type Theory a surprisingly strong expressive power, equal to that of classical higher-order logic. Note, however, that the only formulas we can *prove* are those of constructive mathematics: the logic of Type Theory is *intuitionistic* which means that we cannot rely on classical axioms such as *excluded middle* ($A \vee \neg A$) or *double negation elimination* ($\neg\neg A \Rightarrow A$).

When it comes to specifications of programs, this is not a bug, but rather a feature. The requirements on a program can be expressed at the level of types, for example

```
f : (x : X) -> Pre x -> Sigma (y : Y) (Post x y)
```

is the type of a function that takes as input elements of a type `X` having the property `Pre`, and delivers elements of a type Y which are in the relation Post with the input. The Sigma in the return type of f represents a dependent pair: this consists of a value y : Y and of a value of type Post x y, depending both on x and on y.

After a long detour on Type Theory, we can finally answer the question stated at the beginning of this paragraph: "How do we actually verify that a program fulfills a specification?" This is done by implementing another program. For example, a verified implementation of our square-root program R would consist of the function itself

```
R : (x : Double) -> 0 <= x -> Double
```

and of another function

```
sqrtR : (x : Double) -> 0 <= x -> R(x) * R(x) = x
```

Notice that the type of sqrtR (R is a square-root function) encodes the logical proposition (1) with $\mathbb{R}$ replaced by Double. If an implementation of sqrtR can be type-checked to be total (to terminate for every input x : Double and for every evidence that 0 <= x) it is in every respect a proof that R is a square-root function and the equivalence between implementing sqrtR and proving the corresponding logical proposition has been established rigorously (Wadler, 2015).

As already discussed, implementing sqrtR in this form is impossible and the equality R(x) * R(x) = x has to be weakened to equality up to a suitable tolerance.

The example shows that even expressing specifications for (let apart verifying) functions that perform floating point operations is not a trivial task. Indeed, as of yet we cannot rely on an easy-to-use form of validated numerics (Tucker, 2011) - this is still an area of on-going research (see e.g. (Boldo and Melquiond, 2017) to get an idea of the current state of the art).

Perhaps not surprisingly, the approach to program specification and verification based on Type Theory works extremely well for all what is not directly based on floating point computations.

It has been successfully applied in e.g., producing a verified C compiler, CompCert (Leroy, 2009); developing database access libraries which statically guarantee that queries are consistent with the schema of the underlying database (Oury and Swierstra, 2008); implementing secure distributed programming (Swamy et al., 2011); implementing resource-safe programs (Morgenstern and Licata, 2010), (Brady and Hammond, 2012); and many others.

In TiPES WP6, we apply verified methods to provide decision makers with policies that are machine-checked to be optimal. We discuss why in the next section.

Formal methods as a surrogate for empirical evidences

We have argued that as we move away from well understood problems to climate models, integrated assessment models, agent-based models and, more generally, methods for climate policy advice, program verification becomes mandatory.

But, in science and engineering, we have plenty of examples of programs that are not verified and yet are successfully applied to inform policy advice. For instance, deep neural networks are routinely applied for decision making in routing problems, gaming, and medical screening.

Cannot we provide accountable policy advice without having to care about program verification? This is a very legitimate question that needs to be addressed with some care.

Ideally, we would like climate policy advice to be based on empirical evidences. This is not only because our understanding of the climate system is far from being perfect.

Most importantly, we know that optimal decisions in matters of climate policy necessarily depend on uncertainties that we can hardly estimate: how likely is it that decisions about emission reductions taken, say, by the EU, are actually going to be implemented over the next decade? What about decisions taken by China or by the USA?

We all know too well that facts do not always follow decisions and that, more than often, taken decisions are not implemented or are implemented with delays. Legislations have large inertia and governments do not always manage to comply with their own decisions.

We know that these uncertainties, but also uncertainties on the consequences of trespassing critical climate thresholds or on the collateral effects of geo-engineering approaches towards mitigating the impacts of climate change, do have an impact on optimal emission policies.

In other words, we know (for sure because we have obtained these assessments by applying verified methods) that decisions on emissions that are optimal when we assume these uncertainties to be zero become sub-optimal when we account for these uncertainties properly (Botta et al., 2018).

Thus, we would like climate policy advice to be based on empirical evidences. But gathering empirical evidences in matters of climate policy is nearly impossible!

Not even global players like China or the USA can afford to perform large scale, carefully designed social experiments in order to assess the effectiveness of, say, carbon taxation schemes. We cannot test two or three carbon taxation schemes on a couple of EU countries to find out which one would be best to adopt on a larger scale.

In other words, we have to advise decision makers without being able to rely on empirical evidences. This is unfortunate but hardly avoidable. In this situation, the only guarantees that we can provide to decision makers come from verified methods. This is not a peculiarity of policy advice in matters of climate: advising governments on how to auction radio frequencies or internet domains (Caminati et al., 2015) faces similar problems, and similar problems are also encountered in policy advice on matters on epidemics, financial markets and taxation and security. Not surprisingly, these are application domains in which formal methods are routinely applied to provide some form of accountability in absence of empirical evidences. They are not ideal but certainly better than nothing.

The bottom line

The main purpose of this note was to discuss why verified programming and formal methods are at the core of the WP6 approach towards "Understanding and communicating the impacts of Tipping Point uncertainties on accountable policies".

We have pointed out the roles of program specification, testing and verification in program development and argued that Type Theory is a useful approach for verified programming.

But most of what we have discussed remains true if we replace the word *program* with the word *problem*[3]!

[3] Indeed, as early as 1932, Kolmogorov showed in a short paper (Kolmogoroff, 1932) (written in German!) that the rules of intuitionistic logic – the logic of Type Theory – can be interpreted as rules of "problem computation"

Thus, this note also points to the fact that Type Theory can be applied as a methodology for understanding and formulating problems and as a vehicle for communication between computer scientists and scientists from other disciplines.

It hopefully also points to the fact that the main role of computer science is not confined to the execution of arithmetical operations or sending data over networks, but is rather to be found in the formulation of concepts, identification and resolution of ambiguities, and, above all, in making our ideas clear.

References

Abrial, J.-R.: The B-Book: Assigning Programs to Meanings, Cambridge University Press., 1996.

Allen, S., Bickford, M., Constable, R., Eaton, R., Kreitz, C., Lorigo, L. and Moran, E.: Innovations in computational type theory using NuPRL, Journal of Applied Logic, 4(4), 428–469, 2006.

Altenkirch, T.: Naive Type Theory, http://www.cs.nott.ac.uk/~psztxa/mgs-17/notes-mgs17.pdf, 2017.

Bird, R.: Thinking Functionally with Haskell, Cambridge University Press., 2014.

Boldo, S. and Melquiond, G.: Computer Arithmetic and Formal Proofs, ISTE Press - Elsevier., 2017.

Botta, N., Jansson, P. and Ionescu, C.: Contributions to a computational theory of policy advice and avoidability, J. Funct. Program., (27, e23), 2017.

Botta, N., Jansson, P. and Ionescu, C.: The impact of uncertainty on optimal emission policies, Earth System Dynamics, 9(2), 525–542, https://doi.org/10.5194/esd-9-525-2018, 2018.

Botta, N., Brede, N., Crucifix, M., Ionescu, C., Jansson, P. and Martínez, M.: Climate science and climate policy, 2020.

Bowen, J.: Formal Specification and Documentation using Z: A Case Study Approach, International Thomson Computer Press., 1996.

Brady, E.: Type-Driven Development in Idris, Manning Publications Co., 2017.

Brady, E. and Hammond, K.: Resource-safe systems programming with embedded domain specific languages, in International Symposium on Practical Aspects of Declarative Languages, Springer, Berlin, Heidelberg„ 2012.

Caminati, M., Kerber, M., Lange-Bever, C. and Rowat, C.: Sound auction specification and implementation, in EC '15 Proceedings of the 16th ACM Conference on Economics and Computation, edited by T. Roughgarden, M. Feldman, and M. Schwarz, pp. 547–564, Association for Computing Machinery, https://doi.org/10.1145/2764468.2764511„ 2015.

Caminiti, L.: Boost C++ libraries Version 1.73.0, https://www.boost.org/doc/libs/develop/libs/contract/doc/html/index.html, 2020.

Claessen, K. and Hughes, J.: QuickCheck: A Lightweight Tool for Random Testing of Haskell Programs, in Proceedings of the 5th International Conference on Functional Programming, pp. 268–279„ 2000.

Clavel, M., Durán, F., Eker, S., Lincoln, P., Martí-Oliet, N., Meseguer, J. and Talcott, C.: All About Maude - A High-Performance Logical Framework, Springer-Verlag., 2007.

("Aufgabenrechnung" in the original paper).

de Moura, L., Kong, S., Avigad, J., Doorn, F. van and Raumer, J. von: The Lean Theorem Prover (System Description), in Automated Deduction - CADE-25, pp. 378–388, Springer International Publishing, Cham,, 2015.

Floyd, R. W.: Assigning Meaning to Programs, Mathematical Aspects of Computer Science, 19–32, 1967.

Ionescu, C. and Jansson, P.: Testing versus proving in climate impact research, in 18th International Workshop on Types for Proofs and Programs (TYPES 2011), vol. 19, pp. 41–54, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, http://drops.dagstuhl.de/opus/volltexte/2013/3899, 2013.

Ionescu, C., Jansson, P. and Botta, N.: Type Theory as a Framework for Modelling and Programming, in Proceedings of the 8th International Symposium, ISoLA 2018, part I, vol. 11244, pp. 119–133, Springer, https://doi.org/10.1007/978-3-030-03418-4/_8,, 2018.

Jones, C. B.: Systematic Software Development Using VDM, Second., Prentice-Hall., 1990.

Kees Doets and Eijck, J. van: The Haskell Road to Logic, Math and Programming, College Publications., 2004.

Kolmogoroff, A. N.: Zur Deutung der intuitionistischen Logik, Mathematische Zeitschrift https://doi.org/10.1007/BF01186549, 1932.

Leroy, X.: Formal verification of a realistic compiler, Communications of the ACM, 52(7), 107–115, https://doi.org/10.1145/1538788.1538814, 2009.

Lions, J.-L. and others: Ariane 5 Flight 501 Failure, Report by the Inquiry Board. https://esamultimedia.esa.int/docs/esa-x-1819eng.pdf, 1996.

Martin-Löf, P.: Intuitionistic Type Theory, Bibliopolis, Napoli., 1984.

Meyer, B.: Design by Contract, Technical Report TR-EI-12/CO, Interactive Software Engineering Inc., 1986.

Meyer, B.: Applying 'design by contract', Computer, 25(10), 40–51, https://doi.org/10.1109/2.161279, 1992.

Meyer, B.: Object-Oriented Software Construction (2nd Ed.), Prentice-Hall, Inc., USA., 1997.

Meyer, B.: Getting a program right, in nine episodes, https://bertrandmeyer.com/category/design-by-contract/, 2020.

Morgan, C.: Programming from specifications, Second., Prentice-Hall., 1990.

Morgenstern, J. and Licata, D.: Security-Typed Programming within Dependently Typed Programming, in Proceedings of the 15th ACM SIGPLAN International Conference on Functional Programming, ICFP, vol. 45, pp. 169–180, https://doi.org/10.1145/1863543.1863569,, 2010.

Moschovakis, J.: Intuitionistic Logic, in The Stanford Encyclopedia of Philosophy, edited by E. N. Zalta, https://plato.stanford.edu/archives/win2018/entries/logic-intuitionistic/; Metaphysics Research Lab, Stanford University,, 2018.

Norell, U.: Towards a practical programming language based on dependent type theory, PhD thesis, Chalmers University of Technology; Citeseer https://research.chalmers.se/en/publication/46311, 2007.

Oury, N. and Swierstra, W.: The power of Pi, in Proceedings of the 13th ACM SIGPLAN International Conference on Functional Programming, https://doi.org/10.1145/1411204.1411213,, 2008.

Plato, J. von: The Development of Proof Theory, in The Stanford Encyclopedia of Philosophy, edited by E. N. Zalta, https://plato.stanford.edu/archives/win2018/entries/proof-theory-development/; Metaphysics Research Lab, Stanford University,, 2018.

RosettaCode: Assertions in design by contract, https://rosettacode.org/wiki/Assertions_in_design_by_contract, 2020.

Swamy, N., Chen, J., Fournet, C., Strub, P.-Y., Bhargavan, K. and Yang, J.: Secure distributed programming with value-dependent types, in Proceedings of the 16th ACM SIGPLAN International Conference on Functional Programming, https://doi.org/10.1145/2034773.2034811,, 2011.

The Coq Development Team: The Coq Proof Assistant, version 8.11.0, https://doi.org/10.5281/ZENODO.1003420, 2020.

Tucker, W.: Validated Numerics: A Short Introduction to Rigorous Computations, Princeton University Press. http://www.jstor.org/stable/j.ctvcm4g18, 2011.

Wadler, P.: Propositions as Types, Commun. ACM, 58(12), 75–84, https://doi.org/10.1145/2699407, 2015.

Webster, M. D.: The Curious Role of "Learning" in Climate Policy: Should We Wait for More Data?, MIT Joint Program on the Science; Policy of Global Change, Report No. 67., 2000.

Webster, M. D.: Incorporating Path Dependency into Decision-Analytic Methods: An Application to Global Climate-Change Policy, Decision Analysis, 5(2), 60–75, 2008.

Wikipedia: Boeing 737 MAX groundings, https://en.wikipedia.org/wiki/Boeing_737_MAX_groundings, 2020a.

Wikipedia: Cold fusion, https://en.wikipedia.org/wiki/Cold_fusion, 2020b.

Wikipedia: QuickCheck - from Wikipedia, the free encyclopedia, https://en.wikipedia.org/wiki/QuickCheck, 2020c.

Decision theory and climate policy

The TiPES-WP6 team

2021-07-09

Climate science contributes to decision making in matters of climate policy by providing estimates of the possible impacts of human activities (mainly due to greenhouse gas emissions but also large scale exploitation of land and water resources) on the climate and, the other way round, of the possible impacts of climate change on societies.

Possible impacts (either of human activities on the climate or the other way round) are called scenarios and the main methodology for generating scenarios in climate science is developing computer-based models and performing model simulations.

While climate models can, up to a certain extent, be validated on the basis of indirect observations of past climates (paleo-climatology) and of a growing amount of direct observations and the (conditional) probabilities of different climate change scenarios (for given anthropogenic forcings) can be estimated, assessing the feedback of climate change on societies is more controversial and requires more interdisciplinary efforts.

Because of this asymmetry, climate science has been so far incapable of providing advice on matters of climate policy that is accountable: decision makers do not precisely know what kind of guarantees they can expect from implementing the advice received.

Integrated assessment models (IAMs) of climate change of the kind discussed in (Nordhaus 2018) have been successfully applied to inform decision making but they have also been criticized, mainly because of three reasons: 1) their lack of predictive capability; 2) their reliance on cost-benefit analysis and marginality assumptions and 3) their focus on deterministic sequential decision problems.

While it is true that relevant IAM outputs (for example a social price of carbon) critically depend on the values of model parameters (climate sensitivity, discount factors) that can hardly be estimated reliably (Pindyck 2017), the lack of predictive capability is not the main reason of concern: IAMs are finally not applied for predicting the impacts of climate change on societies but rather for comparing and understanding such impacts.

The reliance of IAMs on marginality (Sharpe and Nijsse 2021) assumptions, cost-benefit analysis and deterministic sequential decision problems, however, are more concerning limitations: cost-benefit analysis requires monetizing the possible impacts of climate change on societies but we do not know how to fairly price these impacts. And the fact that we also do not know how to reliably attach probabilities to impacts of climate change should not be a justification for falsely assuming them to be known with certainty!

More realistic approaches towards rationalizing climate decisions on GHG emissions attempt at accounting for multi-objective notions of optimality (Carlino et al. 2020) in optimal control or for the impact of uncertainties, e.g. on the inertia of legislations and the capability of (global!) decision makers to actually implement decisions (Botta, Jansson, and Ionescu 2018) and on solar radiation management options (Moreno-Cruz and Keith 2012), (Helwegen et al. 2019).

While these approaches can help understanding how uncertainty and the attitude of decision makers towards uncertainties (risk neutral, risk averse, etc.) affect "best" global decisions and can also help clarify the trade-offs made when defining what is "best", they do not tackle the problem of how independent decision makers in a competitive environment can actually coordinate and agree to implement such decisions.

Answering this question requires integrating, among others, optimal control theory, game theory, political science, climate economics and formal methods and is a relatively new research area (Heitzig, Lessmann, and Zou 2011), (Heitzig 2012).

In a nutshell: decision theory (or, better, decision theories) play a crucial role for accountable, pragmatic decision making in matters of climate policies.

At this point, however, their most valuable contribution is perhaps to make clear (to policy advisors, decision makers and, by large, to the civil society), the assumptions implicit in rationalizing decision making in matters of climate policy and the fact that best decisions crucially depend on what we are set to achieve on which time scale, on our attitude with respect to risk and on uncertainties that we can hardly estimate, let apart avoid.

References

Botta, Nicola, Patrik Jansson, and Cezar Ionescu. 2018. "The Impact of Uncertainty on Optimal Emission Policies." *Earth System Dynamics* 9 (2): 525–42. https://doi.org/10.5194/esd-9-525-2018.

Carlino, Angelo, Matteo Giuliani, Massimo Tavoni, and Andrea Castelletti. 2020. "Multi-Objective Optimal Control of a Simple Stochastic Climate-Economy Model." *IFAC-PapersOnLine* 53 (2): 16593–8. https://doi.org/https://doi.org/10.1016/j.ifacol.2020.12.786.

Heitzig, Jobst. 2012. "Bottom-Up Strategic Linking of Carbon Markets: Which Climate Coalitions Would Farsighted Players Form?" *SSRN Environmental Economics eJournal.* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2119219.

Heitzig, Jobst, Kai Lessmann, and Yong Zou. 2011. "Self-Enforcing Strategies to Deter Free-Riding in the Climate Change Mitigation Game and Other Repeated Public Good Games." *Proceedings of the National Academy of Sciences* 108 (38): 15739–44. https://doi.org/10.1073/pnas.1106265108.

Helwegen, Koen G., Claudia E. Wieners, Jason E. Frank, and Henk A. Dijkstra. 2019. "Complementing CO_2 Emission Reduction by Solar Radiation Management Might Strongly Enhance Future Welfare." *Earth System Dynamics* 10 (3): 453–72. https://doi.org/10.5194/esd-10-453-2019.

Moreno-Cruz, Juan, and David Keith. 2012. "Climate Policy Under Uncertainty: A Case for Geoengineering." *Climatic Change.* https://doi.org/10.1007/s10584-012-0487-4.

Nordhaus, William. 2018. "Evolution of Modeling of the Economics of Global Warming: Changes in the DICE Model, 1992–2017." *Climatic Change* 149 (4): 623–40. https://doi.org/10.1007/s10584-

018-2218-y.

Pindyck, Robert S. 2017. “The Use and Misuse of Models for Climate Policy.” Review of Environmental Economics and Policy. https://doi.org/10.1093/reep/rew012.

Sharpe, Mercure, S., and F. J. M. M. Nijsse. 2021. “Deciding How to Decide: Risk-Opportunity Analysis as a Generalisation of Cost-Benefit Analysis.” UCL Institute for Innovation; Public Purpose, Working Paper Series (IIPP WP 2021/03).