# Efficient avoidance of tipping points

## A viable cost of carbon

Pierce Donovan[†]

December 2020

Abstract

The increased likelihood of breaching several potential climate tipping points is directly attributable to anthropogenic activity. Beyond these temperature and carbon-based thresholds, we commit the Earth to drastically altered climate dynamics and truly immeasurable damages. In this paper, I develop an emissions scenario that limits the likelihood of crossing an ambiguous tipping point to a specified level at minimal cost. Instead of relegating threshold avoidance to a constraint on an integrated assessment model, I shift focus toward the maximization of a physical measure of [reversible] global warming potential. This change addresses common concerns about the specification of the economic modules in integrated assessment modeling. I also use a modern climate module consistent with the Earth's climate dynamics and carbon cycle and showcase a new joint-chance constrained approximate dynamic programming algorithm that can handle a large and continuous state space.

***Keywords:*** viable control, approximate dynamic programming, shadow value viability, tipping points, carbon pricing, climate change, pollution control, risk and irreversibility
(JEL codes: C6, Q3, Q5)

[†]Department of Economics, Colgate University
**email**: pdonovan@colgate.edu **web**: piercedonovan.github.io
Please note: this is very much a work in progress and I wouldn't recommend citing my results section just yet. I appreciate any feedback you may have for me, so please reach out with questions/comments.

# 1 Rethinking the optimal control of emissions

The 2016 Paris Agreement codified a desire to keep the increase in the global average temperature within the current century well below 2°C above pre-industrial levels and pursue efforts to limit this increase to 1.5°C (Rogelj et al., 2016). Crossing these key thresholds implies an irreversible commitment to intensified warming beyond some acceptable, adaptable amount (Heutel et al., 2016).[1] These targets reflect an understanding that further warming will cause dangerous interference with the climate system.

Where do these thresholds come from? In part, they are politically-motivated. Pushing scientific and political support to coalesce around a particular concrete number—no matter how arbitrary—increases the chances of organization, cooperation, and action. But there is considerable evidence that breaching the above thresholds commits us to a remarkable level of danger. Global warming has substantially increased the probability of crossing various tipping points—climate thresholds that mark critical, irreversible changes to the dynamics of the climate system (Lontzek et al., 2015; Lenton et al., 2019).[2]

It is well-understood that current commitments are insufficient to meet either temperature goal (IPCC, 2018), and thus limit the risk of tipping. The current "Intended Nationally Determined Contributions," a set of voluntary engagements, imply a median warming of 2.6–3.1°C by 2100 (Rogelj et al., 2016). The remaining carbon budget accordant with the higher 2°C target will require near-zero emissions by the end of the century, and if the 1.5°C target is to be taken seriously, rapid decarbonization is required (Dietz et al., 2018).

There is an ever-increasing number of ways for humanity to reduce its carbon footprint. The energy, manufacturing, agricultural, and transportation sectors have seen large technological improvements that decouple the relationship between economic activity and emissions. Human behavior can also be targeted to eliminate unnecessary waste. Regulation, through taxation or subsidy, command and control actions, or cap-and-trade schemes all nudge (or force) the creators of emissions into compliance with public goals. This paper proposes using a shadow value linked to changes in the likelihood of a tipping event to inform the stringency of these regulations. I'll show how the implied loss from an irreversible change to the climate translates to a social damage measure—which I call the viable cost of carbon (VCC)—that can be used to set the strength of environmental policy.

[1] Aiming for lower intermediate temperatures determines how bad things will be in the long-run. For example, decreasing surface albedo (reflectivity) presents a positive feedback loop. As polar ice melts, the Earth absorbs more energy from the sun. This warms the Earth further, so more ice melts, and more quickly.

[2] Unprecedented hot and wet events such as heatwaves, heavy rainfall, floods, droughts, and extremely low Arctic sea ice provide additional motivation (Diffenbaugh, 2020), and the arrival of these disasters can be modeled similarly to that of the structural changes from tipping.

## 1.1 Integrated assessment modeling and the social cost of carbon

Culturally and scientifically, we are foremost concerned with avoiding extreme and unprecedented changes to the Earth's climate—a threshold avoidance problem. But leading economic approach toward emissions control—integrated assessment modeling—does not center analysis around threshold avoidance and instead sets out to maximize the present value of the flow of societal welfare by controlling all of global economic activity (Nordhaus, 1975). Threshold avoidance is often relegated to a constraint layered on top of this ambitious, all-environing objective.

Integrated assessment models (IAMs) provide clear and coherent linkages between key dynamic variables of climate and the economy (Nordhaus, 1975; Pindyck, 2017). Economic activity generates emissions that contribute to warming of the atmosphere and ocean, sea level rise, and changing precipitation patterns. These changes affect the productivity (and welfare) of later generations, which is weighed against enhanced benefits today. Implicit in this formulation is the assertion that future damages are measurable and predictable.[3]

While these models are clear and coherent, the number of assumed elements is huge. The lion's share of concern is with the relatively "heavy-handed" (compared to other modeling efforts in natural resource economics) economic component of IAMs, not the representation of climate dynamics.[4] IAMs agglomerate the many damages due to large increases in global temperatures, but they are commonly criticized for placing too much significance on unknowable structural inputs concerning inter-generational utility, damage functions that map warming to GDP loss, and the scale and scope of industries and countries included (Pindyck, 2013, 2017; Pezzey, 2019).

Critiques like Pindyck (2017) claim that at best, model components may be derived from our [tenuous (Kolstad and Moore, 2020)] understanding of relatively short term studies on the current effects of climate change, but this knowledge loses validity quickly as we extrapolate far out-of-sample.[5] At worst, IAMs are built upon a series of dangerously subjective and non-falsifiable beliefs (Pezzey, 2019). We should at the very least continue to question the validity of the "optimal emissions" strategies prescribed by these models.

[3] Increasingly comprehensive efforts like Cai and Lontzek (2019) (i.e. Cai et al., 2012) have added complexity and applicability, conditional on this very strong assumption.

[4] Dietz et al. (2020) have recently shown that the dynamics in IAMs fail to represent that of Earth's climate—and with great consequence. To reflect this, I use a carbon cycle model (FAIR, Millar et al., 2017) coupled to a global energy balance model (Geoffroy et al., 2013) to satisfy these new concerns as well.

[5] And it should be noted that the results of these studies themselves are often noisy. "Physical changes in the climate due to greenhouse gas emissions are now well-documented, and future changes due to unmitigated greenhouse gas emissions are generally well understood. But quantifying the economic consequences of changes in temperature, rainfall, sea-level, or other climate variables has long been recognized as extremely challenging" (Kolstad and Moore, 2020).

In IAMs, the output of focus and leading policy instrument is the social cost of carbon (SCC)—an estimate of the present value of all the future damages resulting from an additional 1-ton increase in $CO_2$ emissions today, relative to some reference emissions trajectory. Its magnitude is sensitive to the modeling choices above. Depending on the current political will, designated inputs—and thus the SCC—may change significantly.[6] But good tools can be mishandled and this is not solely a problem with the IAM approach. Yet with so many unknowable inputs it isn't clear that these models can ever be tuned to reality, even when in the hands of a "responsible" economist.[7]

It isn't obvious that "optimal emissions" should be guided by the current heavy-handed approach.[8] Instead, an abatement policy can be thought of as a form of insurance where society pays to decrease the likelihood of a [low-probability] catastrophe (Pindyck, 2013). Further, Pezzey (2019) suggests that any carbon prices used to inform climate policies should be based on marginal abatement costs, found by modeling low-cost pathways to physical climate targets.

I combine the two ideas above. A lightweight alternative to integrated assessment is to maximize a metric of [reversible] global warming potential (energy absorbed by the Earth) while constraining the likelihood of irreversible climate change to some maximal tolerable level. This objective minimizes assumed structure by sourcing a measure of benefits from the climate science literature and matches the spirit of the Paris Agreement goals. The "value function" being maximized yields our threshold avoidance budget—the expected amount of remaining warming potential that doesn't induce tipping. This work applies the necessary economic thinking regarding the tradeoffs between contemporary benefits of emissions and the likelihood of future disaster, in a purely physical (rather than behavioral) framework.

Is this "viability" method any better? The manipulation of the horizon over which we aim to succeed or the level of risk we deem tolerable will result in a similar effect on the VCC that changing a damage function would have on the SCC. But the key difference is that it will remain exceedingly difficult for economics to converge on the "correct" specification for inter-generational utility, risk aversion, GDP projections, damage functions, and the scale of industries and peoples considered—while there is plenty of scientific, political, and cultural support backing up the avoidance of dangerous thresholds with some

[6]The SCC dropped from a central estimate of $52 to at most $8.70 (2020 dollars per ton released in 2020) in the transition from the Obama administration to the Trump administration (Auffhammer, 2018).

[7]From Pezzey (2019): "It is probably less scary to believe that damage valuation estimates can usefully advise us, than to accept how deeply uncertain we must remain about the dangers of the one-off, uncontrolled experiment that humanity is conducting on Earth."

[8]To its credit, the SCC promotes thinking about systematic damage valuation and helps us consider the potential populations, locations, and industries at risk and when they might be affected.

[not impossible to agree on] degree of certainty (Weitzman, 2009; Lemoine and Traeger, 2014; Rogelj et al., 2016; Pindyck, 2017; Dietz et al., 2018; IPCC, 2018; Pezzey, 2019; Lenton et al., 2019; Kolstad and Moore, 2020; Diffenbaugh, 2020; Rudik, 2020).

## 1.2 A viable control solution and related alternatives

Some of the core issues I've raised thus far have been addressed in recent papers with diverse stylistic approaches, through compounding uncertainties (Weitzman, 2009), insuring against catastrophes (Pindyck and Wang, 2013), modeling ambiguous tipping points (Lemoine and Traeger, 2014, 2016) or probabilistic climate thresholds (Held et al., 2009; Fitzpatrick and Kelly, 2017), and misspecification (Rudik, 2020). I'll briefly summarize these contributions and and conclude with how my solution builds on the emergent shadow value viability (SVV) approach (Donovan et al., 2019; Donovan and Springborn, 2020) towards a differently-motivated emissions control policy.

Plenty have acknowledged that low-probability, high-impact (tail) events are important to consider. Weitzman (2009) claimed that society actually has an indefinitely large expected loss from such events. Compounding uncertainties preclude trustworthy (or even finite) bounds on the unprecedented changes to planetary welfare that come from huge increases in average global temperatures. From this, an uncertainty-robust modeling effort *should* be focused on limiting the likelihood of tail events.

In cases where uncertainties are characterized by risk, Pindyck and Wang (2013) estimate the willingness to pay to reduce the likelihood of tail events and their potential damages—i.e. a tax on consumption used to insure against a large negative event—although their approach requires strong economic assumptions not unlike the IAMs approach.

The remaining works provide improvements in integrating uncertainty into IAMs. Held et al. (2009) and Fitzpatrick and Kelly (2017) promote the use of probabilistic stabilization targets, which require that an environmental variable stay beneath the target with a maximum allowable probability.[9] In both cases, this amounted to a *constrained* IAM in which welfare maximization—with its drawbacks—was still the ultimate objective.

Rather than probabilistically-constraining climate variables, Lemoine and Traeger (2014, 2016) focus attention on limiting the likelihood of a potential hazard. A subjective probability of reaching an ambiguous tipping threshold cautions the decision-maker to reduce warming relative to an unmodified integrated assessment framework that does not account for tipping. I use a similar ambiguity strategy in this paper.

[9]These papers use a period-by-period probabilistic constraint rather than a joint-chance constraint, which will not respect irreversible nature of tipping.

Rudik (2020) targets concerns about misspecification in IAMs by incorporating (1) a learning process that acknowledges parametric uncertainty, and (2) robust control, which allows the decision-maker even more distrust in their ability to know the full structure of the climate-economy system. His motivation—that large uncertainties about the economic consequences of additional warming imply the need for an explicit modeling response that captures them—provides similar incentive to investigate a viable control approach as well. Both solution methods are complements in resolving the misspecification issue.

Natural resource management in the presence of potential catastrophe is particularly difficult because the benefits from avoiding most disasters are not well-known (although we usually have an idea of the costs of preemptive action). Following the SVV approach in Donovan et al. (2019) and Donovan and Springborn (2020), I focus on limiting the risk of irreversible climate change to some acceptable level over an extended time horizon, at minimum opportunity cost (in physical terms of global warming potential). This allows a decision-maker to proceed without explicit information on the benefits of preemptive action. As a social planner problem, the method involves identifying the loss from crossing an irreversible climate threshold that drives enough ongoing preventative effort to ensure we do not reach this tipping point with a specified level of confidence. With this loss I construct a function that tracks the implied value of the system as it moves towards or away from a potential tipping event.

SVV relieves what is "optimal" from dependence on assumptions about economic parameters or damage functions and places focus on avoiding the losses from irreversible climate change. The marginal present expected loss due to a future crossing of a climate threshold yields the VCC, an alternative social damage measure to the SCC. It captures the trade-off between the marginal user cost of emissions and changes in the likelihood of tipping. Like the SCC, this figure can be used in benefit-cost analysis of competing emissions control programs. The next section outlines this alternative approach.

# 2 A shadow value-based integrated assessment approach

This section provides an illustration of how to frame an emissions control problem as a viable control objective. I will build up the full approach in focused, digestible stages.

In Section 2.1, I summarize the problem with the current representation of climate dynamics in IAMs and the alternative that I use in this paper, which reproduces the temperature and carbon cycle responses seen in the more realistic Earth system climate models used by IPCC. As I take this solution off-the-shelf from the recent climate literature, I relegate a complete explanation of the climate dynamics to Appendices A.1 and A.2.

Section 2.2 stylizes uncertainty about a potential tipping event in the Knightian sense. Rather than assuming we can measure (and respond to) the exact risk of crossing a tipping threshold, I take an ambiguity-style view which makes use of a subjective hazard function.

Section 2.3 applies the SVV approach to the emissions control problem conditional on the structure laid out in Sections 2.1 and 2.2 above. This places the economic focus on the ongoing avoidance of a threshold with a desired level of confidence.

Lastly, Section 2.4 provides a short explanation of an approximate dynamic programming (ADP) algorithm for solving the SVV problem. Appendix A.3 provides a complete detailing of the SVV-ADP algorithm.

## 2.1 A short summary of recent improvements in climate models

The social cost of carbon is determined by perturbing a future emissions scenario with a 1-ton $CO_2$ "pulse emission" in present-day conditions and calculating the present-valued damages associated with this shock. For this to be believable, the climate modules in IAMs should be able to reproduce the response to a sudden increase (pulse) in emissions seen in the more complicated computational Earth system models used for the oft-cited IPCC projections. None of them can, and it matters (Dietz et al., 2020).

Essentially, two key behaviors are awry in the economic models: (1) the delay between emissions and warming is too long and (2) carbon cycle feedbacks are working in the wrong direction (Dietz et al., 2020). Due to the former, "optimal emissions trajectories" allow temperatures to rise too much, since warming happens far into the future. This places extra weight on the choice of discount rate since benefits and damages are realized at very different times. To summarize the latter issue, climate science shows that as carbon sinks become more saturated, their efficacy in removing carbon from the atmosphere decreases; yet in the IAMs, the opposite happens (Joos et al., 2013; Millar et al., 2017).[10] Both of these effects lead to carbon prices that are too low (Dietz et al., 2020).

These issues call for a tune-up of IAM climate dynamics and apply to my potential alternative as well—which requires a reliable metric of global warming potential. The climate module in this paper follows the Finite Amplitude Impulse Response model, or FAIR (Millar et al., 2017)—which incorporates the Geoffroy et al. (2013) energy balance model with state-dependent carbon cycle feedbacks.[11] This coupled energy balance-carbon cycle model introduces the nuance needed to reproduce the behaviour of Earth system models.

The model is structured as follows. First, emissions enter the atmosphere and are

[10] IAMs assume carbon is reabsorbed in proportion to the atmospheric concentration of $CO_2$, but increasing temperatures and carbon previously absorbed both slow the rate of absorption in real carbon sinks.

[11] FAIR was used by the IPCC in its Special Report on Global Warming of 1.5°C (IPCC, 2018).

binned into one of four "active" carbon reservoirs.[12] These four reservoirs will over time donate carbon to various carbon sinks; these absorptions operate on varying timescales. As the sinks become saturated and as the temperature increases, these timescales increase, causing carbon to linger for longer periods of time in the atmosphere (and absorb more energy each additional year the stock lingers).

The sum of the active carbon concentrations determines the current "radiative forcing," which is the level of an energy flow continually absorbed by the Earth system. Consistent with the law of conservation of energy, the Earth must increase in temperature if it is absorbing more energy per unit time. As objects—including the Earth—heat up, they give off radiation of their own, which provides an opportunity for equilibrium. For example, if the level of forcing remains constant, the Earth will heat up until the radiation from the Earth equals the amount it continually absorbs. Temperature is linked to forcing via two coupled thermal reservoirs, one representing the atmosphere, land, and upper-ocean, and the other capturing the deep ocean. Importantly, having two "boxes" allows the system to have a fast temperature response and a slow one, which mimics the behavior of more complicated climate models.

Since I am more or less taking FAIR off-the-shelf, I relegate a complete discussion and derivation of the dynamics to Appendices A.1 and A.2, and leave the dynamics in Section 2.3 abstract. Tackling this more realistic structure in a dynamic programming framework requires new methods, which I discuss in Section 2.4.

## 2.2 Ambiguous tipping events

I follow Lemoine and Traeger (2014, 2016) in assuming the exact risk associated with tipping [as a function of observable variables] is unknown, and characterize uncertainty in the Knightian sense. The subjective hazard probability, Equation 1, can be used by a decision maker unable to quantify the risk of tipping:

$$H(Tip_{t+1}|X_{t+1}, X_t) = \max\left\{0, \frac{\min\{X_{t+1}, \overline{X}\} - X_t}{\overline{X} - X_t}\right\}, \tag{1}$$

where $X_t$ is the current temperature anomaly, $X_{t+1}$ is the projected anomaly conditional on the current action, and $\overline{X}$ denotes the point at which the system almost certainly tips.[13][14]

[12] By active, I mean that they are currently contributing toward warming by absorbing solar radiation, unlike carbon locked up as oil, methane calthrates in permafrost, or carbonates in the ocean.

[13] I have written this hazard function as dependent only on the temperature anomaly, however there are potentially carbon concentration thresholds as well. Most tipping points are classified with respect to temperature thresholds (Lenton et al., 2019).

[14] In practice, I replace $X_t$ with the running maximum temperature achieved by time $t$.

If we observe a higher temperature without tipping, then we know the true threshold must be higher than the current state of the world and the probability density function (of tipping, over temperature anomaly) squeezes into a smaller, warmer domain. This means that the hazard associated with future temperature anomaly $X_{t+1} = x$ increases as the current state $(X_t)$ becomes warmer.

From Lenton et al. (2019), the likelihood of crossing several different tipping thresholds (at any temperature/carbon concentration) is potentially much higher than initially thought, and upper bounds for some of these tipping points may be as low as 3°C. For illustrative purposes, my model considers a single tipping point with an upper bound at $\overline{X} = 5$°C, which roughly corresponds with a subjective probability of 21% for tipping by 2°C, given current temperatures.

## 2.3 Shadow value viability and emissions control

This section shows how to frame the emissions control problem as one of ongoing probabilistic threshold avoidance using the SVV approach—which I'll call SVV-FAIR. A full presentation of the SVV technique can be found in Donovan and Springborn (2020).

Our optimal control objective is to maximize expected present benefits by way of emissions, net the expected present damages due to tipping. In SVV, this is akin to solving for the fixed point of Equation 2 subject to three constraints,

$$
\begin{aligned}
V^{\Omega}(S_t) = \max_{E_t} \left\{ \pi(E_t, S_t) + \beta \cdot \mathbb{E}_{\varepsilon} \left[ V^{\Omega}(S_{t+1}) \mid E_t, S_t \right] \right\} \qquad (2) \\
\text{s.t. } S_{t+1} = G(E_t, S_t, \varepsilon_t) \\
V^{\Omega}(Tip_t = 1) = \Omega < 0 \\
\Pr\{Tip_{t+T} = 0 \mid A^{\Omega}(S_t)\} \geq \Delta.
\end{aligned}
$$

$S_t = \{Tip_t, CE_t, X_{A,t}, X_{O,t}, R_{n,t}\}$, $(n = 1...4)$ is our collection of state variables in the system that evolves according to $G(\cdot)$. Of our climate variables, $CE$ is the cumulative level of emissions since pre-industrial levels, $X_A$ and $X_O$ are the mean global temperature anomalies (relative to 1880) in the atmosphere/land/upper ocean and deep ocean, respectively, and the $R_n$ states pertain to the carbon concentration anomalies in our four atmospheric carbon reservoirs. Each of these state variables—save for $Tip$—evolves deterministically.

Benefits are captured by [absolute] *global warming potential* (GWP) (Aamaas et al., 2013), rather than with a dollar metric which would require additional behavioral assumptions. GWP captures the total amount of energy absorbed by the Earth over some time horizon from an increase in emissions (and thus radiative forcing) today. As it takes time for

emissions to be absorbed into carbon sinks, carbon emitted today will not only increase contemporary warming, but contribute additional warming in the intermediate future as well, as the increased forcing due to additional carbon stock is a flow. The amount of energy absorbed in the current year is given by the current radiative forcing $F(E_t, S_t)$—detailed in Appendix A.2—multiplied by the yearly timestep ($\pi(E_t, S_t) = F(E_t, S_t) \cdot 1yr$), which results in a measure of energy (per unit area, $W \cdot yr/m^2$).

The net present value of the stream of benefits provides the threshold avoidance budget—the expected amount of warming we may allow without crossing a tipping threshold—conditional on the current state. The nature of this metric precludes discounting in a behavioral "time-value" sense. Instead, we could assume that humans have an increasing capacity to mitigate forcing over time, which has a similar effect. This prevents the value function from growing without bound over an infinite horizon.[15]

$Tip$ tells us whether or not the system has reached a representative tipping point. The likelihood of tipping in the next period, given by $H(\cdot)$ in the previous section, provides us with the probability that $\varepsilon$ (Bernoulli) pushes $Tip$ from 0 to 1, where it remains thereafter. Our value function, $V^{\Omega}$, collapses to a single value once $Tip = 1$,

$$V^{\Omega}(Tip_t = 1) = \Omega < 0. \tag{3}$$

$V^{\Omega}(\cdot)$ is pinned to a hypothetical loss $\Omega$ that represents the sudden inability to capture the benefits associated with $\Omega$-units of of additional warming once the system tips. Regardless of the current level of prosperity, we enter a regime of irreversible warming whose avoidance was supposed to be our primary goal. This small modification to the value function further yields a state-dependent expected present loss (Donovan and Springborn, 2020)—the shadow value of tipping:

$$\omega(S_t) = \Omega \cdot \sum_{u=t}^{\infty} \beta^{u-t} \cdot \Pr\left(Tip_u = 1 \mid A^{\Omega}, S_t\right). \tag{4}$$

The loss $\Omega$ propagates to all other states in expected present value form via $\omega(S_t)$, which discounts $\Omega$ by both the time to and likelihood of tipping. $\partial\omega(S_t)/\partial S_t$ provides us with a shadow value in the usual marginal sense, with respect to a change in the [discounted] likelihood. It is specifically $\partial\omega(S_t)/\partial CE_t$ that will provide a viable cost of carbon that captures the marginal increase in the expected present losses from tipping due to a 1-ton increase in present-day $CO_2$ emissions. As the social cost of carbon changes over

[15]The constraints on the value function also preclude an infinity-result, eventually forcing zero—or even negative—emissions for very degraded states and halting the rise in expected cumulative GWP. Because of this, I am able to set $\beta = 1$ in this application and the value function still converges.

time, so will the VCC, which can be determined for any particular state of the world $S$.

To inform the VCC, we must first determine how to estimate $\Omega$. SVV "chooses" the smallest (magnitude) loss that incentivizes enough abatement effort in order to meet a *viability constraint*, which requires the climate system makes it through year $T$ without tipping with at least $\Delta$-% likelihood,[16]

$$\Pr\{Tip_{t+T} = 0 \mid A^{\Omega}(S_t)\} \geq \Delta.^{17} \tag{5}$$

The pair $\{T, \Delta\}$ reflects a social preference, and is subject to some of the same criticisms of IAMs. A longer horizon or a higher confidence level will both increase $\Omega$. Comparatively, the scope and scale of what IAM assumptions must cover is far more egregious and far less transparent. Still, to suggest a potential solution, $T$ and $\Delta$ could be set using the variability in IPCC climate model projections; we can peg our time horizon to the interval over which IPCC projections remain under a tolerable variance, and that variability can be used to set our confidence level as well. In the current model, $T$ and $\Delta$ correspond to the year 2100 (80 years) and 80%, respectively.

This solution admits a policy that efficiently scales as the state varies from near to far from tipping. As the state degrades (likelihood of tipping increases), the VCC will increase to put additional pressure on emissions control. This increase is anticipated and the social planner will aim to avoid an [increasing] marginal cost. The resulting program will endogenously place the climate system safely away from tipping; just "how far" will depend on the speed of the dynamic system, the opportunity costs of abatement, and the preferences given by the viability horizon and confidence level.

## 2.4 An Approximate Dynamic Programming solution method

To solve the continuous-space, many-state model (Equation 2), I use *Approximate Dynamic Programming* (ADP)[18]—a simulation-based method for solving Markov decision processes (Powell, 2011; Springborn and Faig, 2019). SVV-FAIR contributes seven continu-

[16] We are continually-concerned with the next $T$ periods into the future. If the viability horizon did not roll forward with time, by $t+T-1$ our social planner would only be concerned with tipping for one last year, which is inconsistent with avoiding irreversible warming forever. Additionally, an infinite-horizon viability constraint is impractical since there is always a non-zero likelihood of tipping in each period.

[17] This constraint cannot be met everywhere, nor can every starting position be evaluated for feasibility in a multivariate and continuous state space. In SVV, a *viability kernel* is computed—which tells us which starting states are viable—and then we apply this constraint to all states in the kernel. In Appendix A.3, I provide a simpler alternative that is only concerned with the actual region of the state space we find ourselves in today, and a short discussion of the consequences of this choice.

[18] The "approximate" label refers to how the expected value step in dynamic programming is performed; it is not evaluated directly, but estimated using a sample average of a number of stochastic simulations.

ous state variables (and one binary), creating an intractable computational problem using more common solution methods. While methods like value function iteration can be used to solve complex dynamic problems, they suffer from the curse[s] of dimensionality as state or decision spaces increase in size. ADP preserves the continuous nature of the state variables and allows us to solve the problem without excessive computational resources.

The method shifts the accounting of actions so that we track the dynamics of the *pre*-shock state $N_t$ instead of the post-shock state $S_t = f(N_t, \varepsilon_t)$. Writing our value function over the pre-shock state allows us to switch the maximization and expectation operators of Equation 2—and a way around costly numerical integration.

By sampling a starting sub-region of the $N$-space and simulating dynamics forward in time, we can build a synthetic data set of $\{N, V(N|\varepsilon)\}$ observations. Then, regressing $V(N|\varepsilon)$ on $N$, we can estimate $\mathbb{E}_\varepsilon[V(N)]$. Repeated application of these two steps adjusts the value function estimate until convergence. The ADP method and SVV-FAIR implementation are fully explained in Appendix A.3.

# 3 Optimal warming and threshold avoidance

Dynamic programming methods like SVV provide solutions as a function of the state, rather than time, unlike the standard outputs in integrated assessment models. To compare with the existing literature, I generate the expected paths of $CO_2$-equivalent emissions, atmospheric carbon concentration, and surface temperature anomaly under SVV-FAIR, starting from present-day conditions. I then discuss the physical nature of the viable cost of carbon and how it may complement its dollarized cousin.

## 3.1 Optimal warming outcomes

Figure 1 provides visualizations of the expected intermediate-term emissions, carbon concentrations, and surface temperature anomaly under SVV-FAIR—given the chosen parameterization of $\Delta$, $T$, and $\overline{X}$ (80%, 80 years, and 5°C, respectively). Emissions are rapidly reduced; by 2080, SVV-FAIR prescribes negative emissions. Peak carbon concentrations (470 ppm) occur in the 2040's, at which point carbon re-uptake into sinks outpaces decreasing emissions. Not long after, surface temperatures stop increasing, as the evanescent portion of the lagged effect from past emissions dies off. As this expected temperature (1.6°C) causes a system tipping event within 80 years at most 20% of the time, SVV-FAIR does not require a reduction in temperature after its peak—this dominates the remainder of the emissions schedule. By 2100, emissions, carbon concentrations, and surface temper-

ature anomaly are nearly-stabilized at their long-run equilibrium values under the policy (2.7 $GtCO_2$, 442 ppm, and 1.6°C, respectively).

Where does this prescribed scenario fall within the climate literature? The Representative Concentration Pathways (RCPs) are commonly-referenced sets of land use, atmospheric emissions and concentration scenarios consistent with the current climate literature (van Vuuren et al., 2011a). The RCPs differ in assumptions about increasing energy demand, rising fossil-fuel prices and climate policy. Given their prominence, I use them as a reference point for my prescriptive model.

The expected scenario under the emissions policy from SVV-FAIR—given the chosen parameterization of $\Delta$, $T$, and $\overline{X}$—is most similar to RCP2.6 (van Vuuren et al., 2011a,b). The RCP2.6 concentration pathway summarizes the low end of potential concentration

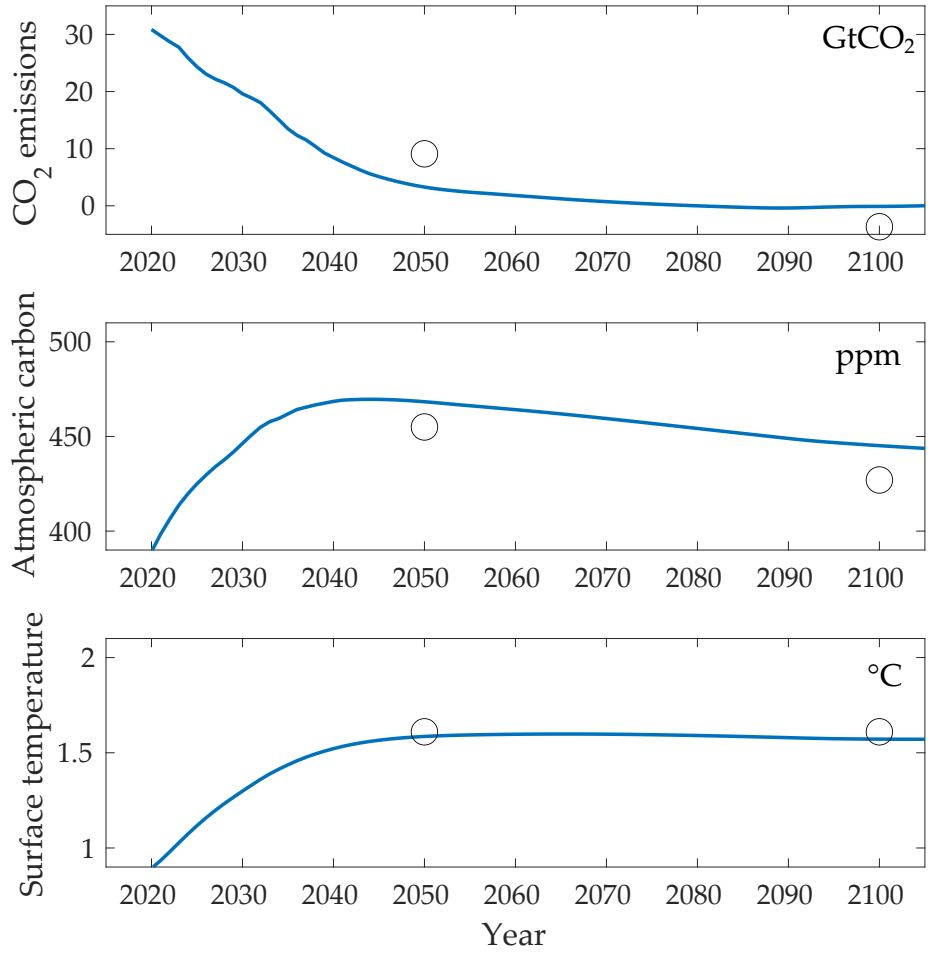


**Figure 1:** Expected paths of $CO_2$-equivalent emissions (gigatons $CO_2$), atmospheric carbon concentration (parts per million by volume), and surface temperature anomaly (degrees Celsius), given the system has not tipped, under the optimal emissions policy. The reference circles represent the central estimates of quantities under RCP2.6 (van Vuuren et al., 2011b) in 2050 and 2100.

trajectories—which aim to limit surface temperature anomaly to 2°C (van Vuuren et al., 2011b). These are technically feasible but very stringent decarbonization scenarios. "2.6" refers to the level of climate forcing in 2100, a value which SVV-FAIR also yields under the current parameterization. The circles on Figure 1 provide reference points for the central estimates of each variable in RCP2.6 in 2050 and 2100.

SVV-FAIR adds to the literature supporting RCP2.6 as the only scenario consistent with avoiding tipping points (Lenton et al., 2019). Lenton et al. (2019) note that the remaining emissions budget for a 50:50 chance of staying within 1.5°C of warming is about 500 $GtCO_2$. SVV-FAIR uses 570 $GtCO_2$ to achieve its chosen goal of 1.6°C.

Recent work by Dietz et al. (2020) provides another convenient model comparison. Their model *DICE-FAIR-Geoffroy* uses an identical representation of the climate and carbon cycle as applied here; the only difference between the models is the economic module. They utilize DICE2016, which I have replaced with the simplified benefits function (in terms of global warming potential), the potential loss due to tipping ($\Omega$), and the viability constraint. To make my comparison, I use the same starting conditions as their paper.

Dietz et al. (2020) have two main model pathways of interest, one that provides optimal emissions that maximise social welfare and another policy that limits warming to 2°C at minimum discounted abatement cost. In the welfare maximization case, their model prescribes a much slower taper of emissions (17.8 $GtCO_2$ in 2100) and thus a temperature increase to 2.95°C above pre-industrial levels—which has a subjective probability of tipping of 50% by 2100 in SVV-FAIR. Their 2°C cost-minimizing path is much closer to the present one, with a 2100 emissions rate of 0 $GtCO_2$ and a temperature anomaly of 1.79°C.

Since SVV-FAIR isn't concerned with behavioral discounting or monetary costs, it prescribes more rapid decarbonization than a welfare-maximizing model. For example, the 2°C cost-minimizing path using DICE-FAIR-Geoffroy puts off decarbonization for longer and requires a more extreme decoupling in the future. In contrast, SVV-FAIR only has high emissions initially because lower temperatures pose less tipping risk. This is convenient for representing a social planner who would like to give themselves as much flexibility in achieving decarbonization as possible.

Note from the SVV-FAIR program in Section 2.3 that the social planner is not trying to maximize the total *amount* of emissions available, but the *impact* of those emissions. If we changed our objective to the former, the emissions path would become much flatter, as atmospheric carbon re-uptake is more efficient when carbon sinks are less saturated. Not only is this not the goal we care about—as climate forcing is what directly increases the temperature—it would put unnecessary constraints on economic activity in the short-term and fail to incentivize long-term decarbonization.

## 3.2 The social costs of tipping

Since SVV-FAIR does not have a module representing the economy, there is no dollarized output to compare against the SCC. This limitation is not as problematic as some may assume. First, the emissions pathway in the previous section prescribes an evolving cap that could be used in a global cap and trade scheme, just like any other emissions trajectory. And in the case of avoiding a dangerous threshold, the additional inter-temporal flexibility of a carbon tax may not be so appropriate anyway—most would probably prefer some guaranteed level of emissions "escapement."

Second, SVV-FAIR has now been shown to be in agreement with other leading integrated assessment models despite having a completely different objective function. Because of this, any SCC-equivalent from this model would be within the common ranges already reported today. This result could potentially be anticipated, as the viability goal is consistent with the same climate dynamics that initially generated support for the 1.5°C and 2°C targets (IPCC, 2014; IPCC, 2018). In practice, modest changes to $\Delta$, $T$, and $\overline{X}$ did not lead to large prescriptive changes.

Third, the SVV-FAIR model outputs provide complementary information to the SCC which is not often available to report. The viable "cost" manifests as a loss in the threshold avoidance budget. This loss, weighted by the likelihood of tipping, reduces the perceived remaining emissions available to us, and alters the rate of decarbonization.

The penalty for tipping ($\Omega$) is $50 W/m^2$, or a fifth of the total reversible global warming potential recommended by the model. This is equivalent to a potential emissions loss of 190 $GtCO_2$ at a baseline of 400 ppm. For reference, this is the same order of magnitude of the expected loss from permafrost methane emissions, should they be unlocked (Lenton et al., 2019). Alternatively, this is roughly five years of present-day global emissions, without any further attempt at decarbonization.[19]

Thus the desire to avoid a representative tipping point is for the first time quantified. The total social costs of tipping are on the same order of magnitude as several years of global fossil fuel based economic activity. While these initial results are not easily converted to the marginal-dollar level, the [co-]benefits of introducing models like SVV-FAIR into the integrated assessment ensemble are now more obvious. Integration will bring us much closer to fully understanding the damages of unchecked climate change.

[19] Another way to think about the penalty is to generate an equivalent starting point that reflects the magnitude of the expected loss from tipping as if it were a true budget reduction. This new initial condition (with reduced emissions budget) is 41 ppm higher than present-day concentrations of around 400 ppm. To put this in units of temperature anomaly, holding atmospheric carbon concentrations at current levels would result in a steady-state temperature anomaly of 1.4°C. Holding them at 441 ppm gives us 1.6°C.

# 4 Discussion

Demand for an alternative method of determining optimal abatement strategies arises namely due to several strong, non-falsifiable assumptions made in integrated assessment modeling (Pindyck, 2017; Pezzey, 2019). The objective of this paper is to provide an illustration of a potential alternative solution.

My approach is not meant to replace integrated assessment, but to give policymakers an additional tool built upon a completely different approach. The alternative optimization objective presented here provides insight toward abatement planning that doesn't come from small variations on the status quo framework.

Unlike integrated assessment, SVV does not aim to capture all of the damages due to incremental warming, nor does it capture the actual damages from crossing a tipping threshold. The method estimates the expected threshold avoidance budget—the amount of global warming potential remaining if ongoing avoidance of a tipping threshold is to succeed—given the current state of the world, which informs an optimal abatement plan without requiring a complete model of the economy. This work applies the necessary economic thinking regarding the tradeoffs between contemporary benefits of emissions and the likelihood of future disaster, in a purely physical (rather than behavioral) framework.

I lean heavily on the climate change literature for inspiration—at a cost of not immediately providing a dollar-ized value for use in policy. The value function estimated here takes on a physical interpretation—which closely aligns with the goals and preferences of climate scientists—although such a metric provides a muddier signal for policy relative to a dollar-incentive measure. The most satisfying improvements to the SVV approach in the climate change space will come from a careful transition from global warming potential to a measure of monetary costs (without bringing in the elements contested previously).

One last thought to consider is that SVV-FAIR implicitly provides a more equitable global abatement policy. The effects of warming are not distributed equitably, and some locations and populations are much more likely to bear the burden of increased warming than others. A "welfare"-based approach only aims to avoid tragedies with higher potential monetary damages—and neglects areas of lower economic potential, regardless of who may live there. In contrast, SVV weighs *any* damages due to tipping as unacceptable (to the same degree). As climate change is a global problem, this shift away from wanting to avoid diminished economic value to preserving some reasonable standard of living for everyone is a refreshing one.

# References


Aamaas, B., Peters, G. P., and Fuglestvedt, J. S. (2013). Simple emission metrics for climate impacts. *Earth System Dynamics*, 4(1):145–170. Publisher: Copernicus GmbH.

Auffhammer, M. (2018). Quantifying Economic Damages from Climate Change. *Journal of Economic Perspectives*, 32(4):33–52.

Cai, Y., Judd, K. L., and Lontzek, T. S. (2012). DSICE: A Dynamic Stochastic Integrated Model of Climate and Economy. *SSRN Electronic Journal*.

Cai, Y. and Lontzek, T. S. (2019). The Social Cost of Carbon with Economic and Climate Risks. *Journal of Political Economy*, 127(6):2684–2734. Publisher: The University of Chicago Press.

Dietz, S., Bowen, A., Doda, B., Gambhir, A., and Warren, R. (2018). The Economics of 1.5°C Climate Change. *Annual Review of Environment and Resources*, 43(1):455–480. _eprint: https://doi.org/10.1146/annurev-environ-102017-025817.

Dietz, S., van der Ploeg, F., Rezai, A., and Venmans, F. (2020). Are Economists Getting Climate Dynamics Right and Does It Matter? *CESifo Working Paper No. 8122*.

Diffenbaugh, N. S. (2020). Verification of extreme event attribution: Using out-of-sample observations to assess changes in probabilities of unprecedented events. *Science Advances*, 6(12):eaay2368.

Donovan, P., Bair, L. S., Yackulic, C. B., and Springborn, M. R. (2019). Safety in Numbers: Cost-effective Endangered Species Management for Viable Populations. *Land Economics*, 95(3):435–453.

Donovan, P. and Springborn, M. (2020). Balancing conservation and commerce: A shadow value viability approach for governing bycatch. *Working Paper*.

Fitzpatrick, L. G. and Kelly, D. L. (2017). Probabilistic Stabilization Targets. *Journal of the Association of Environmental and Resource Economists*, 4(2):611–657. Publisher: The University of Chicago Press.

Geoffroy, O., Saint-Martin, D., Olivié, D. J. L., Voldoire, A., Bellon, G., and Tytéca, S. (2013). Transient Climate Response in a Two-Layer Energy-Balance Model. Part I: Analytical Solution and Parameter Calibration Using CMIP5 AOGCM Experiments. *Journal of Climate*, 26(6):1841–1857. Publisher: American Meteorological Society.

Gregory, J. M. (2000). Vertical heat transports in the ocean and their effect on time-dependent climate change. *Climate Dynamics*, 16(7):501–515.

Hansen, J. (2005). Efficacy of climate forcings. *Journal of Geophysical Research*, 110(D18).

Hansen, J., Ruedy, R., Sato, M., and Lo, K. (2010). Global Surface Temperature Change. *Reviews of Geophysics*, 48(4).

Held, H., Kriegler, E., Lessmann, K., and Edenhofer, O. (2009). Efficient climate policies under technology and climate uncertainty. *Energy Economics*, 31:S50–S61.

Held, I. M., Winton, M., Takahashi, K., Delworth, T., Zeng, F., and Vallis, G. K. (2010). Probing the Fast and Slow Components of Global Warming by Returning Abruptly to Preindustrial Forcing. *Journal of Climate*, 23(9):2418–2427. Publisher: American Meteorological Society.

Heutel, G., Moreno-Cruz, J., and Shayegh, S. (2016). Climate tipping points and solar geoengineering. *Journal of Economic Behavior & Organization*, 132:19–45.

IPCC (2014). *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland.

IPCC (2018). *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*. [Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield (eds.)]. IPCC, Geneva, Switzerland.

Joos, F., Roth, R., Fuglestvedt, J. S., Peters, G. P., Enting, I. G., von Bloh, W., Brovkin, V., Burke, E. J., Eby, M., Edwards, N. R., Friedrich, T., Frölicher, T. L., Halloran, P. R., Holden, P. B., Jones, C., Kleinen, T., Mackenzie, F. T., Matsumoto, K., Meinshausen, M., Plattner, G.-K., Reisinger, A., Segschneider, J., Shaffer, G., Steinacher, M., Strassmann, K., Tanaka, K., Timmermann, A., and Weaver, A. J. (2013). Carbon dioxide and climate impulse response functions for the computation of greenhouse gas metrics: a multi-model analysis. *Atmospheric Chemistry and Physics*, 13(5):2793–2825.

Judd, K. L. (1998). *Numerical Methods in Economics*. MIT Press.

Kolstad, C. D. and Moore, F. C. (2020). Estimating the Economic Impacts of Climate Change Using Weather Observations. *Review of Environmental Economics and Policy*, 14(1):1–24.

Lemoine, D. and Traeger, C. (2014). Watch Your Step: Optimal Policy in a Tipping Climate. *American Economic Journal: Economic Policy*, 6(1):137–166.

Lemoine, D. and Traeger, C. P. (2016). Ambiguous tipping points. *Journal of Economic Behavior & Organization*, 132:5–18.

Lenton, T. M., Rockström, J., Gaffney, O., Rahmstorf, S., Richardson, K., Steffen, W., and Schellnhuber, H. J. (2019). Climate tipping points — too risky to bet against. *Nature*, 575(7784):592–595. Number: 7784 Publisher: Nature Publishing Group.

Lontzek, T. S., Cai, Y., Judd, K. L., and Lenton, T. M. (2015). Stochastic integrated assessment of climate tipping points indicates the need for strict climate policy. *Nature Climate Change*, 5(5):441–444. Number: 5 Publisher: Nature Publishing Group.

Millar, R. J., Nicholls, Z. R., Friedlingstein, P., and Allen, M. R. (2017). A modified impulse-response representation of the global near-surface air temperature and atmospheric concentration response to carbon dioxide emissions. *Atmospheric Chemistry and Physics*, 17(11):7213–7228. Publisher: Copernicus GmbH.

Nordhaus, W. D. (1975). Can we control carbon dioxide? *IIASA Working Paper, WP75063*.

North, G. R., Cahalan, R. F., and Coakley, J. A. (1981). Energy balance climate models. *Reviews of Geophysics*, 19(1):91–121.

Pezzey, J. C. V. (2019). Why the social cost of carbon will always be disputed. *Wiley Interdisciplinary Reviews: Climate Change*, 10(1):e558.

Pindyck, R. S. (2013). Climate Change Policy: What Do the Models Tell Us? *Journal of Economic Literature*, 51(3):860–872.

Pindyck, R. S. (2017). The Use and Misuse of Models for Climate Policy. *Review of Environmental Economics and Policy*, 11(1):100–114.

Pindyck, R. S. and Wang, N. (2013). The Economic and Policy Consequences of Catastrophes. *American Economic Journal: Economic Policy*, 5(4):306–339.

Powell, W. B. (2011). *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Rogelj, J., den Elzen, M., Höhne, N., Fransen, T., Fekete, H., Winkler, H., Schaeffer, R., Sha, F., Riahi, K., and Meinshausen, M. (2016). Paris Agreement climate proposals need a boost to keep warming well below 2 °C. *Nature*, 534(7609):631–639. Number: 7609 Publisher: Nature Publishing Group.

Rudik, I. (2020). Optimal Climate Policy When Damages Are Unknown. *American Economic Journal: Economic Policy*, 12(2):340–373.

Schwartz, S. E. (2007). Heat capacity, time constant, and sensitivity of Earth's climate system. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 112(D24).

Smith, C. J., Forster, P. M., Allen, M., Leach, N., Millar, R. J., Passerello, G. A., and Regayre, L. A. (2018). FAIR v1.3: a simple emissions-based impulse response and carbon cycle model. *Geoscientific Model Development*, 11(6):2273–2297. Publisher: Copernicus GmbH.

Springborn, M. R. and Faig, A. (2019). Moving Forward: A Simulation-Based Approach for Solving Dynamic Resource Management Problems. *Marine Resource Economics*, 34(3):199–224. Publisher: The University of Chicago Press.

van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G. C., Kram, T., Krey, V., Lamarque, J.-F., Masui, T., Meinshausen, M., Nakicenovic, N., Smith, S. J., and Rose, S. K. (2011a). The representative concentration pathways: an overview. *Climatic Change*, 109(1):5.

van Vuuren, D. P., Stehfest, E., den Elzen, M. G. J., Kram, T., van Vliet, J., Deetman, S., Isaac, M., Klein Goldewijk, K., Hof, A., Mendoza Beltran, A., Oostenrijk, R., and van Ruijven, B. (2011b). RCP2.6: exploring the possibility to keep global mean temperature increase below 2°C. *Climatic Change*, 109(1):95.

Weitzman, M. L. (2009). On Modeling and Interpreting the Economics of Catastrophic Climate Change. *Review of Economics and Statistics*, 91(1):1–19.

# Appendices for "Efficient avoidance of tipping points"

## A.1 A simple climate model using the law of conservation of energy

The Earth is a thermal reservoir that exchanges energy with the rest of the universe by absorbing and emitting energy. This combination of absorptions and emissions is called the Earth's radiative balance. Here I derive a simple "energy balance model" (North et al., 1981) similar to those seen in Gregory (2000), Schwartz (2007), Held et al. (2010), and Geoffroy et al. (2013). This type of model describes the response of Earth's mean temperature anomaly to a change in radiative forcing (any stimulus that affects the flux of energy to and from the Earth). The "two-box" style model here is closest to that of Geoffroy et al. (2013), and shares that model's parameterization.[1]

Our first assumption is that the total heat $H$ (energy transferred per unit area, $W \cdot yr/m^2$) in the Earth climate system is linearly dependent on temperature $T$ (°C) (at least for tiny fluctuations), related by the specific heat of the Earth, $\mu$ ($W \cdot yr/m^2$°C), the amount of energy needed to raise the average temperature of the Earth by 1°C,

$$H = \mu T. \tag{A.1}$$

The amount of heat in the system varies with time $t$, and the net heat flux can be represented by the heat entering and leaving the system,

$$\dot{H}_t = H_t^{in} - H_t^{out}. \tag{A.2}$$

If energy is to be conserved, incoming radiation must be increasing the Earth's temperature. Radiative cooling counteracts this; as the Earth heats up, it radiates outward more intensely. The Stefan-Boltzmann law provides the relationship for an emitting object,

$$H_t^{out} = \sigma T_t^4, \tag{A.3}$$

where $\sigma$ is the Stefan-Boltzmann constant. Expanding Equation A.3 about a baseline temperature $T_0$ yields a new variable $X_t$, called the temperature *anomaly*,

$$\begin{aligned} H_t^{out} &= \sigma T_t^4 = \sigma (T_0 + X_t)^4 \\ &= \sigma T_0^4 \left(1 + \frac{X_t}{T_0}\right)^4 \\ &\approx \sigma T_0^4 \left(1 + 4\frac{X_t}{T_0}\right) \text{ (for small perturbations)} \\ &= H_0^{out} + 4\sigma T_0^3 X_t, \end{aligned} \tag{A.4}$$

[1] This exposition is adapted from my undergraduate physics capstone project (Donovan, 2015).

where the penultimate line employs a binomial approximation. Plugging this into Equation A.2 and noting from Equation A.1 that $\dot{H}_t = \mu\dot{X}_t$, our net flux becomes

$$\dot{H}_t = \mu\dot{X}_t = H_t^{in} - H_0^{out} - 4\sigma T_0^3 X_t. \tag{A.5}$$

By choosing to call $F_t = H_t^{in} - H_0^{out}$ the net forcing ($W/m^2$) at some time $t$ relative to some baseline, and making the substitution $\lambda = 4\sigma T_0^3$, we arrive at our equation of motion for global average temperature anomaly,

$$\mu\dot{X}_t = F_t - \lambda X_t. \tag{A.6}$$

Equation A.6 relates an increase in energy to the system to the mean global temperature response. These forces, e.g. solar radiation, higher greenhouse gas concentrations, or reflection due to surface albedo or aerosols, affect the balance of incoming and outgoing radiation in Earth-atmosphere system (Hansen, 2005). For example, the forcing associated with a doubling of atmospheric $CO_2$ is well-agreed upon, roughly $4W/m^2$ (Hansen et al., 2010). This means that if we were to increase mean atmospheric $CO_2$ concentrations to $800ppm$, we would impose an additional flux of $4W/m^2$ to be continually absorbed by the Earth system, which will heat up until the resulting outward radiation balances the added incoming flux. This results in further warming beyond a single period because the Earth does not instantaneously adjust. This "thermal inertia" is captured by $-\lambda X_t$. $1/\lambda$ denotes the equilibrium climate sensitivity, or the long term temperature rise required to offset an additional $1W/m^2$ forcing. Following the Stefan-Boltzmann law (Equation A.3), it takes more forcing to induce further temperature change at higher temperatures.

Adding a second layer, or "box" that interacts with the first but does not directly take up forcing from anthropogenic activity, is straightforward. The two coupled equations are now

$$\begin{aligned} \mu_A\dot{X}_{A,t} &= F_t - \lambda X_{A,t} - \gamma(X_{A,t} - X_{O,t}) \\ \text{and } \mu_O\dot{X}_{O,t} &= \gamma(X_{A,t} - X_{O,t}), \end{aligned} \tag{A.7}$$

where $\gamma$ captures the rate of heat transfer between the two thermal reservoirs. The upper box, denoted $A$, is typically considered to capture the atmosphere, land, and upper ocean, while $O$ represents a deep ocean layer.

Table A.1 provides parameter values for the model. The use of two boxes provides important flexibility in the temperature response to an increase in forcing (Joos et al., 2013; Dietz et al., 2020), namely it yields two modes of response to forcing: a fast mode with a time constant of 4.1 years and a slow mode that takes significantly longer for adjustment (219 years) (Geoffroy et al., 2013).

These continuous-time equations are coupled to the discrete-time decision process in this paper, which has a one-year timestep. From numerical investigation of these differential equations (and the ones in the next section) and review of other current coupling efforts (e.g. Dietz et al.

**Table A.1:** Table of energy balance parameters from Geoffroy et al. (2013).

| Parameter | Value | Description |
|---|---|---|
| $\mu_A$ | 7.3 W·yr/m$^2$°C | atmosphere/land/upper-ocean heat capacity |
| $\mu_O$ | 106 W·yr/m$^2$°C | deep ocean heat capacity |
| $\lambda$ | 1.13 W/m$^2$°C | radiative feedback parameter |
| $\gamma$ | 0.73 W/m$^2$°C | heat exchange coefficient |

(2020)), it appears that a discrete approximation over one year yields very similar results to the "true" continuous-time process.

## A.2 Full set of equations and parameters for the FAIR model

This section informs the forcing function $F_t$. Ultimately, we'd like this to capture two effects: (1) as the concentration of "unlocked" carbon[2] increases, the associated radiative forcing increases at a decreasing rate (as the absorption of radiation in $CO_2$'s band becomes more saturated) and (2) the efficacy of carbon sinks decreases as they become more saturated (creating a countervailing effect of longer periods of increased forcing). These two effects turn out to be pivotal for capturing the behavior of computational Earth system models. One successful reproduction of the "full" models, called the Finite Amplitude Impulse Response (FAIR) model, is summarized here. To learn more about the importance of these effects and the calibrations of the model parameters (Table A.2), see Aamaas et al. (2013), Joos et al. (2013), and Millar et al. (2017).[3]

We start with four active carbon reservoirs associated with four different geologic processes of varying timescales. The change in carbon concentration anomaly in each pool, $R_n$, is given as

$$\dot{R}_{n,t} = a_n E_t - \frac{R_{n,t}}{\alpha_t \tau_n}, \; n = 1...4, \tag{A.8}$$

where $E_t$ is the annual $CO_2$ emissions, in GtC ($1ppm\ CO_2 = 2.12GtC$), $a_n$ is the fraction of carbon emissions entering each reservoir, and $\tau_n$ is the decay time constant for carbon to slip back into a "locked" state associated with process $n$. The state-dependent factor $\alpha_t$ is derived below. Total atmospheric $CO_2$ concentrations are given by the sum of carbon in these four reservoirs,

$$C_t = C_0 + \sum_{n=1}^{4} R_{n,t}, \tag{A.9}$$

i.e., the concentration anomalies augment $C_0$, the pre-industrial $CO_2$ concentration.

[2]By "unlocked," I mean "active," e.g. carbon that is playing a role in increased absorption of infrared radiation. Examples of carbon that is currently "locked up" are untapped oil, methane calthrates in permafrost, or carbonates in the ocean. The carbon that leaves the model ultimately joins these sinks.

[3]Smith et al. (2018) provides an extension to FAIR that expands model scope to non-$CO_2$ forcings.

The radiative forcing as a result of the total concentration $C_t$ is

$$F_t = \frac{F_{2xCO_2}}{\log 2} \log\left(\frac{C_t}{C_0}\right) + F_{ext,t}\ , \tag{A.10}$$

where $F_{2xCO_2}$ is the forcing associated with $CO_2$ doubling and $F_{ext,t}$ captures non-$CO_2$ forcing. In this paper, all other forcings $F_{ext,t}$ are first converted to $CO_2$-equivalent concentrations, then converted to forcing with the above formula.

The rates of uptake of carbon from these reservoirs is modulated by the state-dependent scaling factor $\alpha_t$. This changes with current carbon concentrations and temperatures via the 100-year integrated impulse response function, $iIRF_{100}$, which is linked to the total amount of energy absorbed by the Earth over 100 years per GtC released today, often called *global warming potential* (Aamaas et al., 2013; Joos et al., 2013). The $iIRF_{100,t}$ is a monotonic (but nonlinear) function of $\alpha_t$,

$$iIRF_{100,t} = \alpha_t \cdot \sum_{n=1}^{4} a_n \tau_n \left[1 - \exp\left(\frac{-100}{\alpha_t \tau_n}\right)\right]. \tag{A.11}$$

In essence, Equation A.11 sums up the yearly incremental contributions to warming over time from an increase in radiative forcing in an initial period (that then decreases over time), as forcing is a flow (the integration of the impulse response function—whose form arises from small perturbations of equations A.8 and A.10—provides each of the four exponential terms in the sum).[4] This metric produces an effective number of years of warming under a non-degrading jump in forcing. When carbon sinks have relatively high efficacy (low $\alpha$), the $iIRF_{100}$ is smaller.

This relationship between $\alpha$ and $iIRF_{100}$ doesn't have an analytical solution, but Dietz et al. (2020) provide a close-fitting approximation,

$$\alpha_t = 0.0107 \cdot \exp(0.0866 \cdot iIRF_{100,t}). \tag{A.12}$$

To link $\alpha_t$ to temperatures and accumulated carbon, the $iIRF_{100,t}$ is first estimated with the following linear relationship,

$$iIRF_{100,t} = r_0 + r_C C_{acc,t} + r_X X_t, \tag{A.13}$$

$$\text{where } C_{acc,t} = \sum_{u=0}^{t} E_u - (C_t - C_0). \tag{A.14}$$

$C_{acc,t}$ is the carbon accumulated in sinks, i.e. all emissions up to point $t$, minus the carbon still in atmosphere. From Equation A.8 and equations A.11 through A.14, we can see that less $CO_2$ is removed for the same increase in forcing against a higher background concentration of $CO_2$.

[4]The units of "time" can be confusing here, but a constant of proportionality is commonly omitted in this literature. The full (un-normalized) units of the $iIRF_{100}$ are W·yr/m$^2$, energy per unit area, or the total amount of energy absorbed from the additional forcing.

**Table A.2:** Table of FAIR carbon cycle parameters from Millar et al. (2017).

| Parameter | Value | Description |
|---|---|---|
| $a_0$ | 0.2173 | geological re-absorption |
| $a_1$ | 0.2240 | deep ocean invasion/equilibration |
| $a_2$ | 0.2824 | biospheric uptake/ocean thermocline invasion |
| $a_3$ | 0.2763 | rapid biospheric uptake/ocean mixed-layer invasion |
| $\tau_0$ | $10^6$ yr | geological re-absorption |
| $\tau_1$ | 394.4 yr | deep ocean invasion/equilibration |
| $\tau_2$ | 36.54 yr | biospheric uptake/ocean thermocline invasion |
| $\tau_3$ | 4.304 yr | rapid biospheric uptake/ocean mixed-layer invasion |
| $r_0$ | 32.40 yr | preindustrial $\text{iIRF}_{100}$ |
| $r_C$ | 0.019 yr/GtC | increase in $\text{iIRF}_{100}$ with cumulative carbon uptake |
| $r_X$ | 4.165 yr/°C | increase in $\text{iIRF}_{100}$ with warming |
| $F_{2xCO_2}$ | 3.51 W/m$^2$ | forcing from a doubling of atmospheric $CO_2$ |
| $C_0$ | 587 GtC | pre-industrial $CO_2$ concentration |

In order to solve this model each timestep, first we calculate the $iIRF_{100,t}$ given the current state of the world and the chosen level of emissions. This lets us solve for $\alpha_t$, which informs the rates of decay of carbon from the four active reservoirs. Given this and the contemporary emissions $E_t$, we know the updated concentration of carbon in the atmosphere which allows us to calculate the level of forcing that will update the temperature in our two thermal reservoirs.

There are seven state variables needed to track the evolution of this model: two temperature anomalies $X_A$ and $X_O$, our four active carbon reservoirs $R_1 ... R_4$, and the cumulative level of emissions $CE_t = \sum_{s=0}^{t} E_s$. The use of Approximate Dynamic Programming, detailed in the next appendix, enables us to couple these complex dynamics to a Markov decision process, which is both novel and groundbreaking in integrated assessment.

## A.3 An Approximate Dynamic Programming solution method

To solve the continuous-space, many-state model (Equation 2), I use *Approximate Dynamic Programming* (ADP)[5]—a simulation-based method for solving Markov decision processes (Powell, 2011; Springborn and Faig, 2019). SVV-FAIR contributes seven continuous state variables (and one binary)—creating an intractable computational problem using contemporary solution methods. While methods like value function iteration can be used to solve complex dynamic problems, they suffer from the curse[s] of dimensionality as state or decision spaces increase in size. ADP preserves the continuous nature of the state variables and allows us to solve the problem without excessive computational resources.

This appendix sets up a generic SVV-ADP algorithm that can be applied to SVV-FAIR in this paper. We start by writing a simplified dynamic programming problem,

$$\begin{aligned} V^s(S_t) &= \max_{A_t} \{\pi(A_t|S_t) + \beta \cdot \mathbb{E}_z[V^s(S_{t+1})\} \\ \text{s.t. } S_{t+1} &= Z_{t+1} \cdot G(A_t, S_t), \\ A_t &\in \mathcal{A}(S_t). \end{aligned}$$

A small modification to our problem facilitates the use of ADP. If we shift our accounting of events by one operation, so that we consider the pre-shock state (or post-decision state, Judd (1998)) $N$, rather than the post-shock state $S = Z \cdot N$, our problem becomes:

$$\begin{aligned} V^n(N_t) &= \mathbb{E}_z\left[\max_{A_t} \{\pi(A_t|N_t) + \beta \cdot V^n(N_{t+1})\}\right] \\ \text{s.t. } N_{t+1} &= G(A_t, Z_t \cdot N_t), \\ A_t &\in \mathcal{A}(Z_t \cdot N_t). \end{aligned}$$

Note that this formulation takes the stochastic shock as given and allows us to switch the maximization and expectation operators.[6] By sampling the $N$-space and simulating a number of draws from the $Z$-distribution, we can build a synthetic data set of $\{N, V(N|Z)\}$ observations. Then, regressing $V(N|Z)$ on $N$, we can estimate the expected value $\mathbb{E}_z[V(N)]$.

In essence, the computation bounces between two components: (1) creating a bank of simulations that sample the $N$-$V$ space, relating the pre-shock state $N$ to a "draw" of the value function $V(N)$ using a current guess and (2) estimating an updated value function via regression. The following solution algorithm below builds upon the one found in Springborn and Faig (2019) in order to solve chance-constrained dynamic programming problems.

[5]The "approximate" label refers to how the expected value step in dynamic programming is performed; it is not evaluated directly, but estimated using a sample average of a number of stochastic simulations.

[6]After solving this shifted problem, we can integrate the value function over $P_z(S|N)$ in order to recover the value as a function of the post-shock state.

**SVV-ADP algorithm**

1. Provide guesses for the loss $\Omega_{j=0}$ and value function $V_{k=0}$; choose the number of simulations $m$, the length of simulation chains $T$ per value function updating step, and the relative weight of new simulation information to the existing guess.
2. For each value function update step, generate a bank of simulations.
   (a) Randomly sample starting points (states $N_{t,m}$) for $m$ simulation chains.[7]
   (b) Generate shocks to be applied to each chain to prep $N_{t+1,m}(N_{t,m}, Z_{t,m}|A_{t,m})$.
   (c) Maximize the value function $V$ at the sampled points $(N_{t,m}, Z_{t,m})$, given the current value function guess $(V_k)$.[8]
   (d) If continuing chain ($t < T$), record the new state, $N_{t+1,m}$, given action $A_{t,m}$. Return to step (b) and draw another set of shocks. At $T$, proceed to (3).[9]
3. Weight $(\delta_k)$ the new value function observations $(V_k(N_t|Z_t))$ against $(1 - \delta_k)$ the existing guess $(V_{k-1}(N_t|Z_t))$, where $k$ is the $k^{\text{th}}$ regression step.[10]
4. Regress the weighted $V_{k,k-1}(N|Z)$ on $N$ using the $m \cdot T$ observations.[11] This computes the expectation step and updates our guess of V.[12]
5. Compare the new guess $V_{k+1}$ against the old guess $V_k$. Using some convergence criteria, return to (2) if the new guess is still sufficiently different.
6. Check the fraction of the most recent simulations that crossed the viability threshold over the desired horizon.[13] Compare to desired confidence level. If too high, choose a smaller $\Omega_{j+1}$ than the current guess and return to (2). Otherwise, increase the loss. Terminate according to some convergence criteria for $\Omega$.[14]

[7]The rule you use for sampling your state space may be specific to your application. You may want to over-sample certain areas, or rule out implausible states. See the next page for sampling in this case.

[8]Maximization can be done using any standard/off-the-shelf method.

[9]Occasionally, long chains have decreasing marginal information per step if the partially-controllable dynamics quickly converge on a sub-region of the state space. This may introduce redundant observations.

[10]This ensures the value function will not wildly jump around and will eventually converge. Initially, we would want to put lots of weight on new information, but for later value function updates, any new simulations are less likely to improve our guess of the value function and so we would decrease their weight.

[11]At our threshold, the value function should pass through [or sufficiently close to] $\Omega_j$. One easy way to generate additional data consistent with both the bank of simulations and this constraint is to take the observations that didn't tip and switch each $Tip_{t,m} = 1$.

[12]A non-parametric fit like Gaussian process regression won't impose a particular shape on the value function, which would (1) affect the optimal policy and (2) bias the stochastic shocks and thus the estimation of V. Additionally, non-parametric methods don't "explode" where observations are scarce or near the bounds of the state space.

[13]We may want to choose to use a larger number of short simulations for the purpose of the regression step, and a smaller number of long simulations to evaluate viability.

[14]There is potential to "double dip" on the simulations in each $k$ step. If the viability constraint is terribly off in step (d), one can short-circuit the inner loop and move on to a different $\Omega$ guess.

In order to visualize jumping from Section 2 or Appendices A.1-A.2 to an actual coded version of SVV-FAIR using ADP, it may be helpful to reference the below list, which provides an order of computational events within the simulation step. There are many moving parts to this dynamic model, and even with all of them now laid out, it turns our that putting the pieces together in a working application is still a challenging puzzle.

Many of the variables are bounded in their sampling; not only does this save significant computational time, but it ensures that the value function estimate is well-fit to the sub-region of the state space that is actually plausible (and relevant). Within the simulation step (2) above, the below list provides an efficient way of sampling the relevant starting states. For each $m$,

1. Set $Tip_{t,m} = 0$, as there's no reason to sample a region with a degenerate value function (once tipped, $V^{\Omega} = \Omega$).
2. Sample cumulative emissions $CE_{t,m}$—which is needed to inform values of the other state variables—from a distribution[15] of potential values between where we are today and an upper bound for where we may be in the near future.[16]
3. Given emissions, we can now draw each $R_{n,t,m}$, from long time constant to short.
    - Set $R_{0,t,m}$ (reservoir with longest time constant) to $a$0-% of $CE_{t,m}$, as this carbon reservoir does not adjust on human timescales.
    - Shorter-timescale geologic processes reach equilibrium faster, which bounds how much lower carbon concentration anomalies in the faster reservoirs can be relative to $R_{0,t,m}$ (e.g. $R_{0,t,m} > R_{1,t,m} > R_{2,t,m} > R_{3,t,m} \geq 0$).
4. Draw $X_{A,t,m}$ from 1.2°C (today) to slightly above $F_t/\lambda$ (we have $F_t$ after summing $R_{n,t,m}$). This represents systems that range from having just seen large increases in emissions to having already equilibrated and are now observing negative forcing.
5. Draw $X_{O,t,m}$, which is plausibly within $(\gamma/\mu_0)/(\lambda/\mu_A)$-% of $X_{A,t,m}$ due to coupling.
6. To simulate chains, we can now draw tipping $\varepsilon_{t,m}$ shocks, and use the now-deterministic dynamics to move forward in time.

[15] Uniform, if you'd like to have an agnostic prior. I use this for all of the sampling here.

[16] This starts us down a slightly different road than the SVV algorithms in Donovan et al. (2019) and Donovan and Springborn (2020), which first identify the region of the state-space in which it is possible to meet the viability constraint—the *viability kernel*—and then tweak $\Omega$ in order to ensure that under the optimal policy, any starting state in the region will be viable. That method is time-consistent, while in the present case, we could potentially find ourselves in a degraded state somewhere down a simulation chain (where continued viability is in jeopardy) and wish we could do more than our feedback policy prescribes. The previous technique is much less feasible given the enormous (many-times infinite) number of initial states we would have to check for feasibility. However, since this is specifically the emissions-control problem, we are much more concerned with having a strong value function fit that reflects where we are today and where we may go in the intermediate future, and not necessarily recovery from incredibly degraded states.