What are we measuring?

Measurement and aggregation issues in economics, with an application to climate risks

Discussion paper

Authors: Eddie Gerba and Gireesh Shrimali

November 2025

Abstract

This paper reviews the twin challenges of measurement and aggregation in economics and the natural sciences, with climate risk as a guiding example. It synthesises a broad range of theoretical and empirical perspectives, tracing ideas from early systems theory to modern macroeconomic debates, and compares the approaches of economics, complexity science, and climate science to the micro–macro aggregation problem. Several key conceptual tensions are highlighted—most notably the "micro–macro gap"—and the limitations of traditional models when confronted with heterogeneity, deep uncertainty, and non-linear feedbacks are demonstrated, especially in the climate-risk context. It also reviews emerging methodologies and proposes integrated frameworks to combine micro-level detail with macro-level consistency. Finally, the paper outlines a roadmap for future research and policy, advocating interdisciplinary collaboration, improved data infrastructure, and adaptive modelling strategies to better capture climate change.

***Keywords**:* Micro-macro gap, open vs closed aggregation, microfoundations, climate risks
***JEL codes**: B41, C18, C80, E10*

Gerba, Eddie. and Shrimali, Gireesh, (2025) "What are we measuring?" Measurement and aggregation issues in economics, with an application to climate risks. Also available in SSRN

Correspondence: eddie.gerba@bankofengland.co.uk

Acknowledgement

Eddie Gerba: Bank of England, London School of Economics, and University of Oxford-Sustainable Finance Lab. **Gireesh Shrimali**: Smith School of Enterprise and Environment, and Sustainable Finance Group at Oxford University.

The work has greatly benefitted from comments and suggestions by Max Huppertz, Junyi Zhao, Lukasz Krebel, Marcin Borsuk, and Oxford-CGFI Fellows. This project was inspired by the thoughtful discussions around the 2023 Hybrid Workshop on Microfoundations in Measurement and Theory.

Discussion paper

1. Introduction

Measurement and aggregation are interlinked challenges at the heart of understanding complex systems in both economics and the natural sciences (Sonnenschein, 1972; 1973; 1982; Simon, 1962). At the most fundamental level, the problem can be framed as: ***How can myriad micro-level elements or actors be meaningfully combined into coherent macro-level quantities or dynamics, without losing essential information?*** (Simon, 1962). This question surfaces in economics as the classic *aggregation problem* – how to derive reliable macroeconomic relationships from individual behaviour – and in fields like ecology or climate science as the problem of *coarse-graining* complex systems (e.g. summarising an ecosystem or climate system's behaviour in tractable form, a term initially adopted in statistical physics by Boltzmann and Gibbs).

Our analytical review explores systematically these issues, from an interdisciplinary as well as inter-methodological angle. That is atypical in the literature and allows us to link theoretical (or conceptual) contributions across disciplines to empirical challenges and practical problems in climate prudential policy. To illustrate, we conceptually contrast closed to open aggregation and examine their implications for climate stress testing. We also discuss the inherent challenges of complexity and uncertainty in climate risk measurement, highlighting important trade-offs in any metrics or composite indicators, and provide a few (conceptually grounded) tentative solutions (e.g. scenario analyses, climate VaR, impact chains, and hierarchical models). We end the paper with some early suggestions for integrated frameworks and show how the proposed tools can be applied to specific policy considerations. We hope to substantially expand on this in subsequent papers.

We use climate risk as a recurring case study, while noting climate-specific nuances along the way. Climate risk – encompassing physical risks from climate impacts and transition risks from the shift to a low-carbon economy – is a domain where measurement and aggregation challenges are notably pronounced. Climate risk involves multi-dimensional, deeply uncertain, long-term processes that strain conventional statistical tools, and it requires combining insights from physics, economics, and other fields. By examining climate risk, we illustrate how general principles play out in practice, and how advances in one field (e.g. complexity theory) might inform another (e.g. macroeconomic stress testing for climate).

We begin the paper with early theoretical work by Herbert Simon on system hierarchies and then chronologically examine subsequent contributions in economics (Section 2). We then turn to empirical applications in Section 3, examining how measurement conventions (like national income accounting in economics or risk metrics in finance) can create gaps between theoretical aggregates and observed data. Section 4 surveys methodological innovations intended to bridge micro and

macro – from agent-based models and network analytics to scenario-based stress tests and "impact chains" in climate risk analysis. In Section 5, we discuss policy implications: why these ostensibly technical issues matter for real-world decision-making in financial regulation, macroeconomic policy, and climate risk management. Finally, Section 6 offers proposed solutions and future directions, suggesting research avenues to overcome current limitations.

2. Theoretical Foundations: From Micro Behaviours to Macro Patterns

2.1 Hierarchical Systems and Near-Decomposability

One of the first theoretical contributions on the aggregation issue came from Herbert Simon's work in the early 1960s, introducing the idea of *nearly decomposable systems*. Simon (1962) observed that many complex systems in nature and society are organised in hierarchies – think of an economy comprising industries, firms, and individuals, or a biological organism comprising organs, tissues, and cells. In such systems, interactions within a subsystem are much stronger or more frequent than interactions between subsystems. Simon showed that if this condition of near-decomposability holds, one can derive powerful simplifications:

- Within subsystem, internal equilibrium is reached in the short run, largely independent of other subsystems' short-run dynamics. For example, each room in a well-insulated house reaches its own temperature equilibrium rapidly, without immediately depending on the temperature of other rooms (Simon, 1962).
- Across subsystems, only aggregate variables matter in the long run. In Simon's example, over longer periods the rooms do affect each other, but only through an aggregate like the total heat flow or average house temperature, not through every microscopic detail of air molecule exchange (Simon, 1962).

Simon concluded that in near-decomposable systems very little information is lost by moving to an aggregate description. The fine-grained interactions across components are so weak (or slow) that they can be neglected or summarised. This provided a theoretical rationale for why macro-level constructs (like total output of a sector, or average temperature of a region) can be meaningful and predictive: if the system is structured right, the macro aggregates obey their own approximate laws.

It's important, however, to note the caveat: not all systems are neatly hierarchical. Simon acknowledged that some systems have strong, global interactions where "each variable is linked with almost equal strength to almost all others" – in such cases, the near-decomposability assumption breaks down (Simon, 1962). Economic systems can at times resemble this non-decomposable case, especially through factors like widespread market sentiment or dense network connections (e.g. a tightly integrated financial network where every institution is connected). In those situations, aggregate behaviour may not simplify.

Still, Simon's insight had a long-lasting influence. It underpins the modular design of many large-scale economic models (treating, say, the household sector and the production sector as separate blocks that interact via only a few summary variables), and it resonates with multi-scale modelling in physics and other fields. In short, if an economy has a near-decomposable structure, one can derive coherent macro behaviour from micro foundations with minimal loss of information. *This was an optimistic message:* it suggested we can build macro-models grounded in micro behaviour, at least in structured scenarios where heterogeneity has limited forms.

However, one should be careful about assuming a given decomposition as "natural". In Simon's example, "room" is already an aggregate of molecules; we choose to group them that way. So different partitions of the same system could either satisfy or violate the conditions. In practice, modelers impose a hierarchical structure (e.g. sectoral models) hoping that cross-group interactions are sufficiently weak.

2.2 Aggregation in Economic Theory – The Microfoundations Challenge

Subsequent economic theory uncovered severe hurdles in aggregating micro behaviours into a neat macro model. In the 1970s and 1980s, as economists pushed for rigorous microfoundations (partly in reaction to the ad-hoc nature of earlier Keynesian models), results emerged that were devastating to simple aggregation. The most famous are the Sonnenschein–Mantel–Debreu (SMD) results (Sonnenschein, 1972; 1973; 1982). Simply put, SMD showed that if you have many heterogeneous consumers, each following standard rational choice, the *market demand function* that results can be almost any shape – i.e. there is no guarantee it obeys the nice downward-sloping "law of demand" unless you impose very restrictive assumptions (such as all consumers having identical, homothetic preferences). In other words, *without strong assumptions, the aggregate of rational micro behaviour need not act like a "rational individual."* The market doesn't behave like any single representative consumer (Sonnenschein, 1972; 1973; 1982). This implies a representative-agent model might be a poor substitute for a truly heterogeneous economy. The consequence is that macroeconomic relationships (like an aggregate demand curve or an aggregate saving function) can't be theoretically derived except in special cases. (Strictly speaking, the SMD results concern the aggregate excess demand function, not directly the market demand curve, but since excess demand = demand minus supply, and in simple cases supply is fixed, it carries over – essentially, they showed that without homogeneity assumptions, microeconomic rationality places almost no restriction on the shape of aggregate excess demand.)

One intuitive way to understand this is that individual idiosyncrasies can cancel out or amplify in unexpected ways when summed. Unless there is some common structure (like everyone reacts identically to price changes, or incomes and preferences are distributed in a specific pattern), the aggregate demand might wiggle up and down with no simple pattern. Thus, aggregation can introduce new degrees of freedom – the macro behaviour might include factors or "pseudo-random"

elements that don't correspond to any single micro behaviour. This result spurred several reactions in economics:

- Some researchers, undeterred, assumed the special cases where exact aggregation is possible. Much of modern macroeconomics (*Real Business Cycle and New Keynesian traditions*) adopted the representative-agent device explicitly. By assuming all individuals are identical (or that their preferences/incomes fit certain conditions like Gorman's form that allow exact aggregation), they closed their models – at the cost of abstracting from heterogeneity (Zhu, 2025). This approach yields a specific type of microfoundation (everyone is optimizing, so it's derived from micro principles), but arguably a hollow one, since it knowingly assumes away the interaction complexities that the SMD theorem warned about. As economist Alan Kirman famously stated, *"We wanted microfoundations to add realism, but the only way to get them without losing coherence is to assume all individuals are the same – which is itself unrealistic."* (Kirman, 1992, paraphrased).
- Other economists turned to heterogeneous-agent models despite their complexity. By the 2000s, advances in computation made it feasible to simulate models with many different agents (households with different wealth, firms with different productivity, etc.). These models (sometimes called Heterogeneous Agents New Keynesian (HANK) models in macroeconomics) revealed, for instance, that aggregate consumption might respond to income shocks quite differently once one accounts for distributional effects. For instance, if some consumers live hand-to-mouth while others save, a fiscal stimulus redistributing income has a bigger effect on total consumption than in a representative-agent model. However, solving these models analytically is hard; one often resorts to numerical methods or approximations (e.g. using a few summary statistics of the distribution as "sufficient statistics" for aggregate behaviour).
- In parallel, outside the mainstream, agent-based computational economics emerged, inspired by complexity science. These agent-based models (ABMs) drop the requirement of finding explicit equations for aggregates. Instead, one simulates each agent and lets macro patterns emerge. For example, an ABM might show that even if each agent follows simple behavioural rules, the aggregate output can exhibit realistic business cycle fluctuations or income distributions with fat tails. These patterns are *computed* rather than derived. The drawback is that they can be hard to interpret or to map to traditional economic variables, but they squarely address the heterogeneity that SMD highlighted: in an ABM, aggregates are outcomes, not assumptions.
- A final group of economists embraced pragmatic empirical approaches that use aggregate data without full micro underpinning (like Vector Autoregressions). This approach does not require taking a strong view on micro behaviour at all; it simply models aggregate and might allow them to be influenced by distribution indirectly (via observed aggregate variables). It's agnostic about micro, focusing on time-series correlations at the macro level.

This debate is not merely academic – it matters for domains like climate economics too. If representing a whole economy's climate damage or mitigation behaviour with "a representative firm" or "a representative consumer" is very misleading when impacts are unevenly distributed (e.g. climate change might devastate some regions and barely affect others – there is no 'average'

region), then macro climate-policy models need to account for heterogeneity or risk erroneous predictions.

It's worth noting that some recent theoretical work tries to salvage aggregation by identifying conditions under which micro–macro consistency can be achieved. For example, the concept of **exact aggregation** requires specific functional forms (Gorman, 1953 showed that if individual demand functions are linear in income, aggregates act nicely – but that's a strong assumption). Essentially, if individuals differ only in ways that can be encapsulated by one or two summary parameters, then an aggregate representative agent can exist. But those conditions are rare.

It's important to remember that individuals maximizing utility doesn't guarantee society maximizes anything coherent, unless preferences or distributions are constrained. This realization forced macroeconomists to either simplify micro diversity or embrace new tools to handle it.

2.3 Measurement and Aggregation in Practice

Alongside theoretical aggregation issues, there is a practical angle: *how to measure real-world aggregates*. Interestingly, official statistical practices sometimes force a kind of aggregation that theoretical models don't account for. Zhu (2025) draws a distinction between **closed aggregation** (the theoretical notion of summing micro variables within a model) and **open aggregation** (the methods national statisticians use to produce macro totals from diverse data).

Consider real GDP. In theory (*a closed-aggregation perspective*), one might define real output as the sum of all firms' outputs adjusted for prices. In practice, statisticians use index-number formulas (chain-weighted indices, Laspeyres/Paasche methods, etc.) to calculate real GDP, ensuring that certain identities hold (e.g. total output = total expenditure = total income in national accounts by construction). This procedure may not coincide with a simple sum of micro production functions. As Zhu points out, *national accounts impose a top-down consistency:* they reconcile data from different sources, apply imputation for missing data, and use conventions (e.g. how to treat owner-occupied housing or R&D) that may have no micro-level counterpart. The result is that macro aggregates like GDP or CPI are somewhat constructs of their own – they have an independent standing and are not purely the sum of micros an economist's model might envision (Zhu, 2025).

One consequence is the so-called **micro–macro gap**: a model might perfectly describe each household's consumption, but when you aggregate it, you might not get the same number as the official GDP consumption, because the latter was computed by a different method (*survey data adjusted to match tax data, etc.*). Conversely, macro models sometimes "chase puzzles" in data that exist only because of how data are aggregated. For example, a well-known discrepancy in the US is that the sum of all household surveys reports less consumption than the national accounts do – partly due to underreporting in surveys and different accounting definitions. A model might try to explain a change in the consumption-to-income ratio that is really an artifact of measurement choices, not an actual behavioural shift (Zhu, 2025).

Zhu's perspective is that to truly reconcile micro and macro, we may sometimes need to adjust how we measure the macro. He suggests creating **satellite accounts** for things like climate damage or intangible investment – in essence, additional accounting systems that capture things standard GDP does not. Alternatively, he suggests forcing models to produce aggregates that match how data are constructed – e.g. building the Consumer Price Index formula into the model's structure so that the model's "inflation" is measured the same way as in data (Zhu, 2025). This is a reversal of the usual approach: *We adjust data to fit models; instead, we should adjust models to fit data definitions.*

The broader point is that aggregation is as much empirical as theoretical – deeply context-dependent. Every discipline has measurement protocols. For climate risk, think of how composite risk indices are built in vulnerability assessments (Fritzsche et al., 2014 – the "Impact Chain" approach used by GIZ); these effectively aggregate underlying factors with certain weights and formulas. These choices can introduce biases or hide variability. For example, the widely used "*global average temperature*" aggregates extremely heterogeneous local temperatures – useful for tracking climate change, but it can mask local extremes. Similarly, a *global climate risk index* might combine economic losses, human fatalities, and ecological damage into one number per country, but that involves (explicitly or implicitly) value judgments about trade-offs between money, lives, and environment (Fleurbaey, 2009; Winsberg, 2012).

Perhaps macro aggregates behave smoothly partly because of how they're measured. Alternatively, our macro data may mislead if we assume they came simply from summing typical agents. In economics, an example is the discrepancy between "consumer price inflation" experienced by households and the official CPI: if consumers substitute towards cheaper goods when one item's price rises, their actual cost-of-living increase is lower than a fixed-basket CPI would indicate. Historically, the Boskin Commission (1996) highlighted that the US CPI was overstating true inflation (and understating real consumption growth) because it didn't account for substitution and quality changes properly. In response, methods changed (chain-weighting, hedonic adjustments). But a macroeconomist unaware of those index formula issues might think consumers had some inexplicable boost in purchasing power.

In national accounting terms, **open aggregation** is the inductive, measurement-based construction under macro constraints, whereas **closed aggregation** is deducing aggregates from a theoretical micro model. They don't always align. To bridge these, economists sometimes incorporate measurement into theory. For example, make your model generate data that, when fed through a realistic national accounts' procedure, yields the patterns seen in actual macro data (*rarely done in full, but conceptually possible*).

Concretely, an important practice is **distributional national accounts** – recent efforts (e.g. by Piketty and others) to reconcile micro data with macro totals. For instance, the US and EU now produce Distributional National Accounts that allocate aggregate GDP or wealth to population

percentiles, ensuring the micro distribution sums to the official macro totals (*Federal Reserve's Distributional Financial Accounts, ECB's Distributional Wealth Accounts, etc.*). This requires adjusting micro data to match the aggregates. It's an example of modifying micro measurement to hit macro constraints.

Another divergence arises in finance: summing up individual firms' self-reported climate risks may not equal the systemic risk measured at a system-wide level. If each firm assesses its risk independently, some risks (like a systemic carbon price shock) might be underestimated (each firm assumes "if I'm in trouble, others aren't necessarily", but in reality, all might be hit together). Conversely, there could be double counting: if a power producer and a manufacturer both count the risk of a carbon price on electricity, and one supplies the other, the same risk is counted twice in a naive sum. The macro stress test would handle that differently.

How we measure things influences aggregate behaviour. Aggregation problems can sometimes be mitigated by better measurement – e.g. collecting more granular data or designing metrics that capture distributions. In climate risk, regulators now ask for exposure metrics like "share of portfolio in flood zones" – a simple aggregate, but one that retains some distributional insight (*it tells us how concentrated risk is*). They complement that with scenario loss metrics for specific event severities. Together, these give a richer picture than a single number (Bank of England, 2022).

However, the more complex the picture, the harder to communicate or use. Hence there is always a pressure to aggregate into key indicators. As we do that, we must remain aware of what's lost. A recurring theme will be the need for multiple perspectives rather than over-relying on any single aggregate measure for complex phenomena.

To ground this discussion, consider Table 1 (in Section 4), which contrasts how economists, complexity scientists, and climate scientists each handle measurement and aggregation. For example, economists historically used representative agents (losing heterogeneity), climate scientists use multi-model scenarios (losing a single clear prediction, but exploring uncertainty), and complexity science uses simulations (losing closed-form solutions but capturing emergent effects). Each approach has trade-offs.

Before moving on, one more empirical point: the role of tail risks. Empirical studies increasingly suggest that focusing only on averages is insufficient – one must account for rare catastrophes. In finance, this led to metrics like Value-at-Risk (the loss at a 99th percentile scenario). In climate economics, analogously, researchers look at worst-case tails (like the low-probability, high-impact scenario). Aggregating tail risks is paradoxical: summing expectations might understate true risk if distributions are very skewed. Often analysts prefer to present separate aggregates for expected loss and for tail loss. For example: "We expect $X million annual loss on average (median scenario), but in a 1-in-100-year event, the loss could be $Y million." Here $X is an aggregate (mean) and $Y is another aggregate (extreme quantile). Both derive from underlying distributions but highlight

different parts of it. The challenge is communicating two numbers instead of one – but it's often necessary to avoid a false sense of security from a single average. In practice, people can handle this: for instance, when you fly, you think "the probability of a crash is ~0, but if it happens it's fatal." You don't average those to conclude a small injury – you accept there are two aspects of the risk. Similarly, presenting both typical and tail outcomes gives a fuller picture.

In conclusion, measurement sets the stage for aggregation. If we measure in a way that respects heterogeneity, our aggregates can be more informative. If we measure in coarse terms, then any further aggregation will compound the loss of information. Climate risk has forced practitioners to adopt richer measurement (scenarios, ranges, multi-metric disclosures) because the problem demands it. Economics is also evolving – e.g. distributional national accounts provide more granularity than headline GDP growth. We will next delve into methodological approaches across disciplines to handle aggregation explicitly.

3. Empirical Applications and Challenges in Climate Risk Measurement

3.1 Multidimensional Risk Measurement

Before focusing on *aggregation* per se, consider measuring a complex risk like climate change. Climate risk is inherently multidimensional: it involves physical hazards (heatwaves, floods, storms), exposures (people or assets in harm's way), and vulnerabilities (sensitivity and adaptive capacity). The IPCC defines risk as a function of those three elements (IPCC, 2014). Measuring climate risk thus means collecting and combining very different kinds of data – from climate models (hazard probabilities) to asset values and demographics (exposure) to engineering or social indicators (vulnerability). There is no single natural "unit" of climate risk; it has to be constructed (e.g. *"expected monetary loss per year"* as one metric, or an index value on some arbitrary scale).

In practice, organisations often reduce this complexity by creating indices or scorecards. For instance, a bank might have a climate risk score for each sector or loan, rating it High/Medium/Low (NGFS, 2022). This is a form of aggregation at the measurement stage: multiple variables (e.g. flood risk, emissions intensity, disaster preparedness of a counterparty) might be aggregated into one score. While useful for summarising, it hides detail – two firms could both be "High risk" for very different reasons (one faces high physical risk, another high transition policy risk). This echoes our earlier point: the more we aggregate into one number, the more we obscure.

In climate risk, forward-looking approaches have been emphasised. Traditional financial risk models (like value-at-risk) struggle because climate change breaks their assumptions – the distribution of losses is shifting over time and has fat tails (extreme events dominate averages). Regulators and experts note that *multiple scenarios* and *granular data* are needed (FSB, 2021). Instead of a single expected outcome, institutions consider scenario analysis: e.g. a 2°C warming scenario vs a 4°C warming scenario by 2100 and assess risks under each. This yields not one number but a range of

outcomes (Bank of England, 2021; Zhu, 2025). Aggregating this is tricky: do we take an average of scenario losses? a worst-case percentile? There's no objectively correct way; weighting scenarios is subjective. In practice, many report scenario-specific results (e.g. "In Scenario A, expected loss = X; in Scenario B = Y"). This is effectively acknowledging that one aggregate risk number may not capture the situation – context (which scenario) matters.

Even for a given scenario, model uncertainty is large. Different catastrophe risk models can give very different loss estimates for the same event. A recent study by GARP found that vendor models varied widely – e.g. a portfolio's projected loss in a "100-year flood" ranged by a factor of three across models (Paisley and Nelson, 2025). If a bank simply *averages* these, they get a number that no single model directly supports – a form of aggregation (averaging expert opinions) that introduces its own assumptions. Increasingly, firms report a range or ensemble of model outputs rather than relying on a single one.

The challenge then is *presenting this complexity*. Many choose to show distributions or at least quantiles rather than just expected values. This ties back to our discussion of tail risks: for climate risk, one might report "Our annual expected loss is $100m, but in the 95th percentile bad year it could be $300m." While that is two numbers, it conveys both typical and extreme outcomes, which is crucial for prudent risk management.

3.2 Long Horizons and Deep Uncertainty

Climate risk unfolds over very long horizons (decades to centuries) with deep uncertainty – we cannot even agree on probability distributions for different outcomes far ahead. As a result, measurement moves into the realm of scenarios rather than forecasts. We mentioned scenarios above; here we emphasize why they are necessary. If we tried to assign a single probability to, say, 4°C warming by 2100 vs 2°C, it would be hugely disputed. Instead, scenario analysis treats them as conditional what-ifs. This yields multiple conditional aggregates rather than one unconditional aggregate.

For policymakers and planners, this is a communication challenge: how to summarise "climate risk" into a single indicator when it depends on human actions and deep future uncertainties? The answer is often: you can't and shouldn't. Instead, one uses stress test frameworks that acknowledge multiple possibilities. In the Bank of England's 2021 exploratory exercise, for example, banks had to report results under different scenarios (early policy action vs late action vs no action), and the regulator looked at the system's resilience under each. There wasn't one bottom-line number like in a capital stress test; rather, it was a range of outcomes and a qualitative assessment of vulnerabilities.

This multi-scenario approach is essentially *opening up* the aggregation – not collapsing across scenarios but keeping them separate. It's an interesting case where, as mentioned earlier, providing

a dashboard of indicators (one per scenario, plus perhaps a subjective judgment of plausibility) is more informative than any single composite metric.

3.3 Micro to Macro in Climate Economics

Let's illustrate measurement vs aggregation with a concrete task: estimating the impact of climate change on country-level GDP by 2050.

One approach (common in *Integrated Assessment Models*, IAMs) is **top-down**: use an empirical macro damage function (perhaps estimated from historical climate–economy correlations) that directly maps global temperature increase to GDP loss. This gives an aggregate answer but arguably bypasses micro detail (and likely understates extremes, since it's usually an average-effect estimate).

Another approach is **bottom-up**: model impacts on various sectors or regions and sum them. For instance, estimate losses in agriculture, plus losses from sea-level rise on coastal property, plus energy costs, etc., and add them up. This could give a richer picture (some sectors might benefit from mild warming even as others suffer). However, when summed, one must be careful about interdependencies: if agriculture suffers, that affects manufacturing that relies on agricultural inputs, etc. – simply summing sector losses might double-count or omit knock-on effects.

Empirically, bottom-up and top-down methods can diverge significantly, which recalls our earlier point on *incommensurability*: they are effectively different paradigms yielding different aggregates. Studies have found that summing sectoral IAM results can yield a different global damage estimate than using a single aggregate IAM with an overarching damage function. This is analogous to the discrepancy between aggregated micro consumption and national accounts consumption in economics – a sign that something doesn't line up, prompting investigation (see comment by Junyi about "methodological incommensurability").

In practice, both approaches are used and each has proponents. Top-down gives simple metrics for policy (e.g. "X% of GDP by 2100" for given warming), which can be convenient but possibly misleading. Bottom-up can provide insight into which areas are hardest hit, but summing them to a grand total may involve a lot of uncertainty (and often bottom-up analyses come out with larger impacts, because they capture more compounding effects – sometimes leading to scepticism that they might double-count some aspects).

3.4 National Accounting vs Reality

Just as in economics we saw differences between survey totals and national accounts, in climate risk there's an analogous multi-scale accounting problem. The sum of individual firms' reported climate risks might not equal the economy-wide risk for several reasons:

- Systemic interactions: If one sector's collapse cascades through others, each firm on its own might not foresee the indirect impacts that a macro analysis would. For example, Bank A assumes it can sell assets in a stress without moving the market, and Bank B assumes the same – but if both try to sell, prices crash more severely, affecting both. Individually, each measured its risk as moderate; collectively, the risk is high.
- Double-counting: As noted, if multiple entities account for the same underlying exposure (e.g. a power plant's emissions risk might appear on the books of the utility company and the fuel supplier and various investors), a naive sum would overstate total risk. National accounts have to net out inter-company transactions; similarly, aggregating climate exposures requires care to net out linked exposures.

In the NGFS 2022 exercise, results were published both at firm level and system level, highlighting that while individual banks might look okay, certain correlated assumptions meant the system had vulnerabilities if all banks acted similarly (e.g. all banks assumed they could shed carbon-intensive assets – but obviously if all try, who's buying?).

3.5 Key Empirical Message

How we measure variables influences what relationships we observe at macro level. Aggregation problems can often be mitigated by better measurement – e.g., collecting more granular or comprehensive data (so we're not missing chunks that get imputed), or designing metrics that include distribution info (like reporting not just a single risk score but also concentration measures or tail stats). In climate risk measurement, this is evident: regulators ask not just for one aggregate like "climate VaR", but for a set of indicators – e.g. *exposure metrics* (like percentage of portfolio in certain risk categories) and *stress test losses under scenarios*. Together, these provide a mosaic of a bank's risk. If we only had one number, it would either obscure too much or have to be so conservative (to account for tails) that it wouldn't be useful for average conditions.

However, the push to maintain multiple metrics comes at the cost of simplicity. Decision-makers often desire a single rating or capital number. There is thus a temptation to aggregate further (e.g. to combine physical and transition risk into one score, or to reduce a full loss distribution to a single "expected shortfall" figure). That's acceptable if one understands the limitations, but dangerous if that single number is taken as truth. A prudent approach is to maintain a dashboard of key metrics.

We see this multi-metric approach already in macro policy: central banks don't have a single index capturing everything; they look at inflation, unemployment, output gap, etc. Similarly, for climate financial risk, a regulator might monitor: (1) aggregate insured losses to GDP (physical risk indicator), (2) banking sector exposure to carbon-intensive assets (transition risk indicator), (3) tail climate VaR of major portfolios, and perhaps (4) some qualitative preparedness index. Each is an aggregate but captures a different angle. The combination gives a holistic view. This is analogous to how medical doctors look at multiple vitals (blood pressure, heart rate, cholesterol) rather than one composite health score.

Multiple perspectives are needed for complex risks. One metric cannot capture all dimensions and trying to force one can lead to misinterpretation. This understanding has grown, especially in climate risk management.

4. Methodological Innovations for Bridging Scales and Disciplines

Given the difficulties outlined, researchers have developed various methods in different fields to improve how we aggregate information. Here we compare and contrast some key methodological innovations in economics, complexity science, and climate science that address measurement and aggregation challenges. The aim is to see what each discipline can learn from the others, and how, in tackling a problem like climate risk, a hybrid of these methods might be most effective.

Table 1 provides a high-level comparison across a few dimensions (model type, treatment of heterogeneity, treatment of non-linearity/tails, data focus, conceptual tensions, emerging solutions) for three stylised approaches:

1. **General equilibrium approaches** (e.g. DSGE and standard metrics like CPI/GDP),
2. **Complexity Science approaches** (e.g. agent-based models and network models),
3. **Climate Science/Risk approaches** (e.g. IAMs and scenario analysis used in climate policy).

This table is not rigid – these fields overlap (economists are now also building ABMs; climate scientists use economic models, etc.) – but it highlights tendencies.

Table 1: Comparison of methodologies and conceptual approaches across disciplines. (*Note: This is a stylised comparison. In practice, boundaries blur – economists use agent-based models, climate scientists use economic models, etc. But it highlights general differences in emphasis.*).

Approaches - Dimension	General equilibrium (e.g. CGE, DSGE)	Complexity/Simulation (e.g. ABM, digital twins)	Climate Science Practice (e.g. scenario analysis)
Micro–Macro model	Representative agent or aggregate equations are common (assume a "typical" agent or use simplified macro relationships), sacrificing heterogeneity for tractability. (Most economic models until recently imposed aggregation methods differing from index-number practices used in data.)	Agent-based models and network simulations explicitly model many diverse agents and their interactions, letting macro properties emerge (no representative agent). There isn't a single closed-form "macro equation" – the model generates aggregate outcomes via simulation.	Integrated Assessment Models (IAMs) often use a top-down representative agent economy; however, impact models and risk assessments increasingly combine multidisciplinary modules (e.g. climate models + sector economic models) to capture differences across sectors/regions. Climate models themselves are aggregated

			at large spatial scales and then downscaled.
Treatment of Heterogeneity	Often assumed away or highly stylized (e.g. all consumers identical) to get closed-form results. Heterogeneity introduced only in special cases (two-agent models, etc.) – otherwise aggregates might behave erratically (per SMD theorem). Recent emerging work on HANK models is adding back some heterogeneity with numerical methods.	Fundamental to the approach: every agent can be different. The challenge of heterogeneity is tackled via computation rather than assumption. Emergent macro patterns (fat-tailed outcomes, cascades) arise naturally from diverse agent behavior. Complexity models embrace richness of types but may need reduction techniques (clustering agents) for interpretation.	Recognised as crucial: climate impacts are uneven, so analyses distinguish by region, sector, or population group. However, many policy models still used (until recently) a global or national average damage function. Newer climate risk frameworks (e.g. stress tests) segment data (by sector, geography) to keep heterogeneity visible. There is also heterogeneity in time: near-term vs long-term risks handled via scenario pathways.
Non-linearity & Tail Risks	Tended to linearise around equilibria for analytical convenience (e.g. linear approximations of models, assuming normal shocks). Extreme events often treated as exogenous "shocks" rather than modelled. As a result, traditional aggregates can severely understate risk of rare disasters. (That said, some econ models do allow non-linear dynamics, but solving them analytically is difficult.)	Embraces non-linearity: models include feedback loops (e.g. network cascades) and can generate power-law distributions of outcomes. Rare but massive events emerge in simulations. Rather than one outcome, an ABM yields a distribution of outcomes which can be examined for tail characteristics. Complexity theory explicitly studies critical thresholds, tipping points, and phase transitions – i.e. non-linear emergent phenomena.	Non-linearity is explicit: damage functions are often non-linear (e.g. losses accelerate with temperature). Tipping points are studied, though hard to quantify. Scenario analysis captures some non-linearity by considering qualitatively different futures. Moreover, use of extreme climate scenarios (like high-emissions RCP 8.5) brings tail-risk scenarios into planning. Still, some official estimates (like IAM-based social cost of carbon) arguably underweight tail risks.
Data & Measurement Focus	Relies on aggregate official data (GDP, CPI, etc.) which are top-down consistent but may mask micro variation. Micro data used separately (e.g. microeconometric studies) but often not integrated into macro models. There is a tradition of creating indices (CPI, etc.) – aggregating baskets into one number – reflecting value judgments (Fisher, 2005). Recently, more focus on using rich micro data to inform macro (e.g. central	Utilises large micro-level datasets when available (e.g. detailed network data, firm-level data). Measurement is often granular: the state of every agent is tracked. To summarise results, relies on statistical analysis of simulation outputs (distributions, moments). Less reliant on official aggregate metrics, more on raw or synthetic data. However, complexity models sometimes face calibration issues – they produce "what ifs" more than precise fits to data.	Combines diverse measurements: physical metrics (temperature, sea level), economic metrics (losses, costs), and composite indices (vulnerability indices). The practice is to present multiple metrics instead of one (e.g. warming in °C, plus % GDP loss, plus specific risk indicators). However, for policy, composite indices (like climate risk rankings or a single "social cost of carbon") are often created, aggregating many factors into one

	banks using big data on heterogeneity).		score. Data gaps are acknowledged (e.g. missing asset-level data), leading to use of proxies and scenario data rather than purely historical data.
Conceptual Tensions	Micro vs macro: need to reconcile individual optimization with aggregate outcomes leads to paradoxes (fallacy of composition). Ontologically, often assumes a "representative" entity that may not exist. Has struggled with incommensurability of different theoretical constructs (national accounts vs micro concepts, as discussed). Also tension between theoretical elegance and empirical realism.	Reductionism vs holism: acknowledges that the whole can be more than sum of parts (emergence). Does not force one equilibrium paradigm – uses computational experiment to explore possibilities. But then faces interpretability issues: how to map complex simulation outcomes to simpler understanding or policy use? Also, results can be sensitive to agent rules chosen – raising questions of validation.	Different disciplines (climate science, economics, sociology) each have their own metrics and models – integrating them leads to incommensurability problems (e.g. economic cost vs human lives vs biodiversity loss). Often resolved by converting everything to monetary terms (for cost-benefit analysis), which is philosophically contentious. There's tension between short-term measurable risk vs long-term systemic risk (e.g. insurers focus on near-term, climate models on long-term), leading to an aggregation across time that discounts or neglects future risk.
Emerging Solutions	Developing heterogeneous-agent models with tractable summary statistics (e.g. using distribution's moments as state variables) to inform policy. Using *satellite accounts* to better align macro data with theory (e.g. separate accounting for natural capital or inequality). Increased use of micro data to validate macro models (e.g. granular data in central bank policy models). Essentially, economics is slowly moving toward embracing more complexity in models, aided by better computation.	Improving algorithms to coarse-grain models (e.g. find clusters of agents that can be treated as one without much error). Using machine learning as surrogate models to approximate ABM outcomes with simpler equations (to allow faster analysis or estimation). Integrating network metrics into policy frameworks (e.g. stress test triggers if network connectivity indicates vulnerability). Complexity science is also engaging with domain-specific data to calibrate ABMs more credibly.	IAMs are becoming more modular and stochastic, incorporating uncertainty explicitly (e.g. using Monte Carlo ensembles). Financial stress-testing frameworks are evolving to require granular data inputs from firms (so regulators can aggregate consistently). Proposals for hybrid modelling: e.g. run an ABM for one part of the economy (power sector) and link to a DSGE model for another part (the rest of economy), marrying detail with theory. Also, greater emphasis on common scenario sets (e.g. NGFS scenarios) so that different institutions' results can be compared apples-to-apples.

Economics and climate science are increasingly moving toward the complexity/heterogeneity end of the spectrum, albeit slowly. For example, central banks now sometimes use agent-based models or at least heterogeneous-agent models to complement their standard models (Bank of England, 2025). Climate IAMs are incorporating probabilistic elements and multiple regions rather than one global aggregate. The fields are learning that to get a comprehensive picture, one might need to handle more complexity and give up on closed-form elegance.

4.1 Complexity Science Approaches

Unlike traditional economics, which often sought a solvable equation for aggregates, complexity science embraces simulation. As discussed, an agent-based model simulates many interacting agents (each with potentially different rules or parameters) and then computes the emerging aggregate outcomes. This has two clear advantages for aggregation: it preserves heterogeneity (no need to assume agents are identical) and it can capture non-linear interactions (since you literally model the interactions).

For instance, an ABM of an economy under climate stress could model each firm's supply chain; it might show that if a few key supplier firms fail (due to a disaster), the ripple effects cause a non-linear drop in GDP – something a top-down model might miss. Studies have indeed used ABMs to study climate–economic interactions, finding, for example, that damage propagations can make aggregate losses larger than the sum of direct damages – a purely emergent effect (Lamperti et al., 2019).

ABMs do have downsides: they can be calibrated to match known aggregates, but it's hard to ascertain their accuracy out-of-sample; and interpreting *why* an ABM produced a certain aggregate outcome can be challenging (you may need to analyse the simulation microdata in detail to find the causes, essentially doing computational experiments on the model).

Nonetheless, ABMs are increasingly used by central banks for scenarios. The Bank of England and other central banks have experimented with ABMs for financial networks to see system-wide risk. In climate risk, ABMs of things like energy transition (where thousands of firms invest in green tech or not, and banks finance them or not) can reveal possible paths that an average IAM might overlook. ABMs often produce fat-tailed outcome distributions – which is useful for stress testing (you can directly observe worst-case emergent scenarios).

A complementary complexity approach is network models. These focus on the topology of interactions – e.g. a production network linking industries, or a financial network of banks and borrowers. By analysing networks, one can identify where simple aggregation fails: e.g. find central nodes whose failure would disproportionately impact the whole ("super-spreaders" of risk). Network measures like degree centrality or connectedness serve as aggregated indicators of systemic risk beyond just summing exposures. For instance, two sectors might each have 10% exposure to climate risk individually, but if one sector is a critical supplier to the other, their joint impact could be

worse than 20% because one's disruption amplifies the other's losses. Network analysis helps flag such dependencies.

An interesting innovation combining ABM and network thinking is multi-scale modelling: simulate fine dynamics in one part of a system and use a coarse aggregate representation for another part, then link them. For example, simulate firm-level defaults in an agent-based model for the corporate sector, but feed their aggregate effect into a macro model of employment and demand, then loop back. This is complex but some integrated climate–economy models are heading this way (embedding an agent simulation for the energy sector within a broader macroeconomic model for the rest of the economy).

In summary, complexity science contributes methods to deal with aggregation by brute-force simulation and by explicit modelling of interactions. It thus offers tools to explore scenarios where traditional analytic solutions break down. The challenge is to integrate these insights into decision frameworks that typically prefer simpler models.

4.2 Finance Meets Climate: Scenario Analysis and Stress Testing

As noted, scenario analysis is now a mainstream tool, especially in climate risk. Methodologically, scenario analysis is not about aggregation per se, but it affects aggregation by shifting focus to distributions of outcomes rather than a single expected outcome. Instead of giving one aggregate result, you present multiple conditional aggregates (one per scenario). This sidesteps some uncertainty – you don't commit to one "best estimate" when deep uncertainty reigns, thereby avoiding aggregating across highly disparate possibilities.

Financial regulators now do system-wide stress tests: they gather granular data from banks and insurers, run them through scenario models, and then aggregate results to see if the *system as a whole* is resilient (BoE, 2022). In the 2021 BoE exploratory exercise, for example, each bank estimated its losses under scenarios, and then the BoE added those up and also looked at second-round effects (e.g. if all banks cut lending following losses, what's the macro impact?). This revealed inconsistencies – many banks assumed certain things wouldn't all happen at once, but when aggregated system-wide, they clearly could (like all banks simultaneously trying to offload certain exposures). The exercise forced recognition that micro-prudent behavior can sum to macro-prudential risk (a classic fallacy of composition).

From a methodological viewpoint, stress testing introduces *feedback aggregation*: micro responses are aggregated and then fed back to micros. It's a more iterative aggregation than a static sum. We might call it an *iterative fixed-point approach*: guess macro outcome, adjust micro decisions, recompute macro, iterate to convergence. This is akin to solving for equilibrium in ABMs, but often done manually by scenario analysis rounds.

In climate stress testing, one has to aggregate not just within one institution but across many, possibly using common risk factors. For example, the NGFS scenarios provide common macroeconomic pathways. By using common scenarios, results from different banks can be meaningfully aggregated or compared (apples-to-apples). That standardisation is a sort of pre-aggregation alignment – it ensures that when we sum results, differences aren't due to scenario assumptions but reflect actual exposures.

One key point in stress testing is *communication of aggregated results*. Regulators might report a single number like "total projected loss for the banking system = £X billion under scenario Y," which is an aggregate of aggregates (each bank's aggregate loss). But they often also provide the distribution (like which percent of banks have losses above Z, etc.). This again ties to not relying on one metric.

In summary, scenario analysis and stress testing represent *frameworks* that manage aggregation by splitting the problem: we don't aggregate over states of the world (we consider them separately), but we do aggregate over entities to gauge systemic totals, sometimes re-injecting those totals back to entity level to see second-round effects.

4.3 Climate Risk Innovation: Impact Chains and Hybrid Frameworks

In climate adaptation science, a methodology called impact chains has gained traction (Fritzsche et al., 2014; Zebisch et al., 2021). An impact chain is essentially a causal flow diagram linking climate hazards to intermediate impacts to final outcomes. For example: *Drought frequency* ↑ → *crop yields* ↓ → *farm incomes* ↓ → *loan defaults* ↑ → *bank losses* ↑. By mapping this chain, one can identify points to measure and possibly aggregate along the way. It's a bridge between qualitative and quantitative: experts fill in parts of the chain that aren't purely data-driven. When quantifying, you might attach a distribution or function at each link (like an elasticity of yield to drought). Then, to aggregate up, you propagate through the chain.

This approach is somewhat akin to system dynamics modelling and provides transparency about cause-effect structure. It acknowledges that a one-step aggregate ("drought causes X% GDP loss") may miss nuance, so it breaks down the aggregation into stages that are easier to handle and more linear locally.

Impact chains, and similar factor models, help manage aggregation by introducing structure – not everything is lumped together at once; instead, you aggregate sequentially with clarity about what's combined at each step. In the above chain, you aggregate weather into a regional yield effect (using climate model outputs), then aggregate yields into economic loss (using agricultural models), and so on. At each step, uncertainties can be tracked. This is a modular approach to aggregation, contrasted with a monolithic approach (like one big regression of GDP on temperature). The trade-off is complexity and the need for expert input at each stage.

This approach has been used in vulnerability assessments where data is scarce – experts qualitatively assess links and assign scores. It's essentially a semi-quantitative aggregation. For climate risks, GIZ's *Climate Risk Sourcebook* (2021) uses impact chains to involve stakeholders in identifying key risks and then quantifies them with mixed methods, ensuring local contexts (heterogeneity) are reflected before numbers are rolled up.

4.4 Hierarchical Modelling and State-Space Reduction

Borrowing from control theory and applied math, techniques exist to reduce complex models by finding state-space equivalences. One example is MacKay & Robinson (2018) on Markov chains: they show how to merge microstates into macro-states under certain conditions without losing predictive power. The general idea is to find conditions where micro-states can be clustered such that the system's behaviour (at least in some respects) is unchanged by treating all states in a cluster as identical. This is like finding symmetries or redundancies in the system.

In economics, an analogue would be: can we prove that all consumers of a certain type (e.g. same income and preferences) can be treated as one aggregate consumer? Under what conditions? Researchers have solved special cases – e.g., if individuals have quadratic utility and identical coefficients, their risk-taking can be simply summed. But these are narrow.

One promising area is the use of sufficient statistics as mentioned earlier. Instead of aggregating everything, economists try to derive a small set of aggregate metrics that preserve the influence of heterogeneity. For example, in heterogeneous agent models of consumption, it turns out that one or two moments of the wealth distribution (like the fraction of people with no savings) can be a "sufficient statistic" to predict the marginal propensity to consume in aggregate. Thus, giving a model not just average wealth but also that fraction may allow it to mimic the fully heterogeneous outcome. In climate risk, similarly, maybe the exposure of the top 5% most at-risk assets is a sufficient statistic for tail risk, in addition to the mean exposure.

Hierarchical modelling plays into this: you might have a micro layer, a mezzo layer, and a macro layer, and try to ensure the macro dynamics depend on only a few meso-level summary variables. If you can identify those, you reduce the state-space dramatically. This is conceptually how large agent-based models could interface with policy: by extracting summary indices (e.g. a "financial fragility index" from an ABM that goes into a macro policy rule).

5. Policy Implications: Why Getting Measurement and Aggregation Right Matters

Understanding and improving measurement and aggregation isn't just an academic exercise – it has real consequences for policy and management in both economics and climate-related domains. Here we discuss several areas where these issues play out in policy, and how better approaches can lead to better decisions. We consider implications for:

1. **Financial regulation and systemic risk** management,
2. **Macroeconomic policy** and public investment,
3. **Corporate and portfolio strategy**,
4. **Climate policy** and integrated planning,
5. Overarching issues of **communication** and **trust**,
6. Managing **policy trade-offs**,
7. And **policy coordination** on a global level.

5.1 Financial Regulation and Systemic Risk

In banking and insurance supervision, regulators historically looked at institutions individually – each bank's risk metrics had to be sound. However, there's increasing recognition that even if each firm looks stable in isolation, the system as a whole can be unstable (*the classic fallacy of composition*). This is directly an aggregation issue: how do risk exposures add up across firms, and do they amplify each other?

For example, consider climate risk as a systemic risk: One insurer might be fine with its catastrophe exposure, but if all insurers are heavily exposed to certain regions, a mega-disaster could shake the entire sector. This calls for macroprudential oversight – regulators aggregating exposures across financial firms to detect concentration and interconnection. Indeed, joint scenarios (like those produced by the Network for Greening the Financial System, NGFS) essentially aggregate balance sheet data to see system-wide vulnerabilities (NGFS, 2022).

One implication is regulators pushing for consistent climate risk disclosure. If every bank uses a different climate scenario or model, their numbers can't be aggregated meaningfully. Initiatives like the NGFS scenarios provide a common set of assumptions so that results of multiple banks can be summed or compared without "apples vs oranges" issues. In the future, we may see regulators require banks to hold capital against systemic climate risks – requiring to measure the "aggregate tail risk" across the system, essentially an aggregated Climate-VaR. This could involve taking each bank's loss distribution from a scenario and then compounding them (accounting for correlations due to common factors like a macroeconomic downturn triggered by climate events).

Another regulatory implication is in data infrastructure: supervisors increasingly demand risk data aggregation capabilities from firms (e.g. the Basel BCBS 239 principles on risk data). The reason is that in a stress, a firm (or regulator) must rapidly gauge total exposure to, say, "Gulf Coast flooding" or "carbon-intensive borrowers across the portfolio." If the firm's data are siloed (credit risk separate from market risk, etc.), it can't aggregate in time. So, good aggregation techniques need good underlying data systems.

In summary, for financial stability, the lesson is that microprudential safety doesn't guarantee macro stability – we must measure system-wide aggregates (like total exposure to a risk factor) and cap or manage those. That is now leading to heavy emphasis on consistent measurement standards like the Task Force on Climate-related Financial Disclosures (TCFD).

5.2 Macro Policy and Public Investments

For governments, mis-measuring aggregate economic capacity or risk can be costly. For instance, if climate damages are underrepresented in GDP (because GDP doesn't count loss of environmental services), a government might under-invest in adaptation. The seminal Dasgupta Review (2021) on biodiversity argues that national accounts should be adjusted for nature depletion – effectively an aggregation issue (combining produced capital and natural capital into an aggregate national wealth), because ignoring it gives a false sense of economic security.

Similarly, if policymakers use a single metric like GDP growth to guide stimulus and ignore distribution (who's suffering most), they might miss needed targeted interventions. For climate economics, a lot of argument has been about the social cost of carbon – an aggregate monetary value of climate damage per ton of CO_2. That is a huge aggregation: it compresses myriad impacts over centuries and globally into one price. Yet it's used to set carbon taxes. If that number is wrong (say it ignores tail risks or ecosystem losses because those were hard to monetize), then the carbon tax is mis-calibrated. Some economists have called for dual metrics: monetary cost and a separate indicator for non-monetary damage (like lives or species lost). This is similar to how some central banks, such as the Fed, promote maximum employment while maintaining a stable inflation rate jointly rather than each goal separately. So, a policy implication is to use multiple aggregate indicators for decisions, not rely on one magic number when dealing with complex trade-offs.

Another angle: climate change involves future generations, so how we aggregate costs over time (discounting) is a big policy issue. If you aggregate by heavy discounting, you downplay future losses; choose a lower discount rate, the aggregate present cost of future climate change rises dramatically. This "aggregation across time" is essentially an ethical choice. The Stern Review (2006) famously used a low discount rate and got a high cost of carbon; others used higher rates and got lower costs. The policy decisions (how much to invest now to mitigate or adapt) depend on this aggregation choice.

In public investment decisions, cost-benefit analysis often converts all effects to present value dollars (an aggregate). But as mentioned, many climate effects (loss of life, loss of species, very long-term uncertainties) resist monetization or single-score evaluation. The policy trend is thus to consider multi-criteria decision analysis for climate projects, rather than just one net present value. That means treating aggregation carefully – maybe not aggregating certain dimensions at all, or keeping them separate in the decision framework.

5.3 Corporate and Portfolio Strategy

Companies and investors also aggregate risk within their portfolios or operations. If a firm's own risk assessment aggregates all climate risks into a single "score", it might miss that one part of its business is critically exposed. Best practice emerging is heatmaps that show exposures by category (so aggregation is partial – within categories – but not total). For example, a bank will present to its

board a matrix of climate risk: sectors vs geographies, highlighting high-risk combinations (like real estate loans in coastal Florida). That's an aggregation of loan-level data into a 2D matrix, which retains more info than a single total number.

Investors are increasingly looking at portfolio *greenness* or *brownness*. Indices like portfolio carbon intensity (tons CO_2 per $ revenue aggregated across holdings) are used to align with climate goals. If those aggregates are not measured correctly (company disclosures issues again) or not correlated with actual risk (e.g. carbon intensity doesn't reflect physical risk), they could mislead investment choices. There's active work on improving ESG metrics aggregation – e.g. weighting by ownership share, avoiding double-counting across asset classes – to make sure that when an investor says "our portfolio's carbon footprint is X", X is computed meaningfully and comparably.

One strategic implication for companies is the concept of real options and flexibility. If aggregate metrics are very uncertain, companies might prefer flexible strategies that perform reasonably under multiple aggregated outcomes. For example, if the future carbon price is deeply uncertain (because micro behaviour and policy could lead to very different macro outcomes), a company may invest in technology that is viable in both high and low carbon price scenarios – essentially hedging against the macro aggregate uncertainty. This is a form of robust decision-making acknowledging that aggregated forecasts are highly uncertain.

In risk management, firms also have to decide how to allocate capital or limits to different lines. If a risk metric doesn't fully aggregate risks (due to non-linearity or diversification), they might hold extra buffers. For climate, some banks voluntarily add "management overlays" – e.g. they might limit exposure to certain sectors even if their model (aggregating current data) suggests it's fine, because they foresee that model might not capture future shifts (like a sudden policy change). That's an implicit recognition of aggregation limits.

5.4 Climate Policy and Integrated Planning

At the governmental level of climate policy, measurement and aggregation issues are integral to things like forming a Nationally Determined Contribution (NDC) to emissions reduction. Countries aggregate their various sectoral plans to a single number (e.g. "40% reduction by 2030"), which is a communications necessity. But the danger is the aggregate target might be met on paper while missing key components (like reducing some sectors a lot and others not at all could have different social implications than a uniform reduction). Recognizing this, some countries present multiple targets (emissions plus renewable share plus efficiency improvements) to cover multiple dimensions.

In adaptation policy, governments often use aggregated indices (like vulnerability indices that combine many factors) to allocate resources. If those indices are poorly constructed, money might go to wrong places. For instance, an index that averages exposure and sensitivity could give the same score to a moderately exposed, highly sensitive area and a highly exposed, moderately

sensitive area – but the needed interventions differ. A policymaker needs to peek into the components, not just the aggregate. So, practice is shifting to scorecards rather than single indices, showing sub-components (exposure, sensitivity, adaptive capacity separately, for example).

Another example: the global climate policy discussion often sums up all countries' pledges and says "we're on track for ~2.7°C warming". That aggregate drives global policy pressure ("the world is not yet at 2°C target"). If some countries measure emissions differently or have different baselines, summing is tricky. There's effort under the UNFCCC to standardize GHG accounting to make these aggregates reliable.

Finally, *just transition* policies highlight distribution: it's not enough to hit an aggregate emissions target if it causes concentrated harm to certain communities. So climate plans now often include separate metrics for fairness (e.g. jobs created/lost in certain regions) rather than one aggregate welfare number.

5.5 Communication and Trust

There's also a meta-implication. If stakeholders suspect that aggregated figures (like climate model forecasts) are masking important issues, trust erodes. Being transparent about how aggregates are formed – maybe providing underlying distributions or at least ranges – can improve credibility. For example, the IMF reports fan charts for forecasts in the World Economic Outlook, showing not just a line but a cone of uncertainty. That uncertainty around median implicitly highlights many potential aggregate outcomes. Likewise, the IPCC reports emphasize confidence intervals and multiple scenarios, not just a single best estimate, to communicate uncertainty.

In contrast, when communication fails: early in the climate debate, some communicated climate risk as "the most likely outcome is moderate warming by 2100" without stressing the fat-tail risk of extreme warming. This led some to complacency and later backlash when worst-case was highlighted. Presenting the aggregate (expected warming) alone was misleading; now communication focuses on *"we have X% chance to stay below 2°C, Y% chance of exceeding 3°C, etc."*

For public trust, showing raw data or multiple views behind an aggregate can help. For instance, publishing the list of projects behind a climate finance aggregate can validate that aggregate in the eyes of observers. Hiding methodology can breed suspicion ("garbage in, garbage out").

5.6 Balancing Simplicity and Accuracy in Policy

Policymakers typically want simple metrics to act on (like a debt-to-GDP ratio or a temperature goal). The tension is those simple metrics often hide complex interrelations. There's a push for what some call *dashboard approaches*: instead of one aggregate, a small set of key indicators.

For climate financial risk, as noted, a regulator might monitor: (1) aggregate insured losses to GDP, (2) banking sector exposure to carbon-intensive assets, (3) aggregate climate VaR of portfolios, among others. Each is one number, but collectively they cover different angles. The combination gives a more holistic view.

This multi-metric approach is analogous to how some central banks monitor many sources of risks to assess financial stability (amongst other things, authorities monitor credit, liquidity, market, and solvency risks). It accepts that no single aggregate can capture everything.

In summary, the pursuit of better measurement and aggregation leads to better-informed policy in several ways: it can reveal hidden concentrations of risk that require pre-emptive action; it ensures fairness or effectiveness by not averaging away crucial differences (like ensuring vulnerable communities aren't hidden by average outcomes); and it helps allocate resources efficiently by highlighting where marginal impact is greatest. Conversely, poor aggregation may increase the probability of inaction (e.g. the financial system is stable because each firm looks sound,) or misdirect efforts (spending on average needs while neglecting pockets of extreme need).

Looking forward, we might see collaborative data environments (like public data utilities for climate risk) to ease consistent aggregation. If every bank feeds loan data into a central climate risk database, regulators can pull aggregates on demand. This raises competition and privacy issues, but technically it's feasible and would produce more reliable system aggregates than trying to sum banks' heterogeneously modelled outputs.

5.7 International Policy Coordination

Climate change is global; aggregation issues appear at the international level too. Summing up all countries' climate pledges gives an implied warming (currently around 2.5–3°C). That aggregate drives global policy discussions ("the world is on track for X°C"). If countries measure emissions differently or have different baselines, summing is tricky – hence efforts to standardize GHG accounting via the Paris Agreement's transparency framework. It's analogous to how in economics the IMF standardized national accounts so global GDP or debt can be aggregated reliably.

Policymakers need to be aware of how aggregation can mislead or enlighten. Best practices include:

- **Using multiple complementary aggregate metrics.**
- **Demanding consistency in measurement across units** to allow valid aggregation.
- **Stress-testing aggregates** under various scenarios rather than assuming one outcome.
- **Keeping an eye on distribution even when acting on aggregates** (e.g., supplement aggregate analysis with distributional analysis).

The climate risk challenge has accelerated improvements in these aspects. We can expect cross-fertilization – e.g., techniques from financial risk aggregation being applied to climate vulnerability assessment and vice versa.

6. Towards Integrated Measures and Aggregation – Conclusion and Roadmap

We have seen how measuring and aggregating complex phenomena – such as economic welfare or climate risk – is fraught with challenges, yet crucial for sound decision-making. The literature across economics, complexity science, and climate science reveals both sobering limitations and promising avenues. Lessons from classical economic theory taught us that simply aggregating individual behaviour can produce nonsense unless structure or assumptions tame the problem. Yet, practical needs push us to summarise vast information into actionable insights – we cannot escape aggregation.

The way forward, underscored by recent advances, is to embrace complexity in our measurement and be nuanced in our aggregation. Key insights and takeaways include:

- ✓ **There is no single silver-bullet metric for climate or economic risk.** Instead, a portfolio of indicators is needed. Climate risk managers should consider physical risk, transition risk, tail scenario impacts, etc., separately before forming a composite view. Effective communication will involve conveying uncertainty ranges, not just point estimates.
- ✓ **Heterogeneity and distribution matter enormously.** Averages can mislead when distributions are broad or skewed. Future research should focus on developing better ways to incorporate distributional information into aggregate metrics – e.g. presenting inequality-adjusted aggregates or risk-adjusted aggregates (where a higher dispersion or tail risk inflates the effective aggregate risk measure). In climate risk terms, that might mean weighting metrics not just by mean outcomes but by concentration of risk (e.g. “40% of our exposure is accounted for by the top 10 polluting companies” is a distribution-aware statement).
- ✓ **Non-linear dynamics mean the sum of parts can behave in unexpected ways.** We must design models (and policies) that consider feedback loops across scales. The use of agent-based simulations and networks alongside aggregate models is a promising practice to test the robustness of aggregate predictions. For instance, if an IAM says “X% GDP loss”, but an ABM of firms shows potential collapse of network production beyond that, policymakers should account for that contingency (perhaps via scenario analysis or precautionary buffers).
- ✓ **Improving data quality and consistency is foundational.** Efforts like standardized climate disclosure (e.g. the new ISSB standards), open climate risk databases, and harmonized national accounting for climate impacts will greatly enhance our ability to aggregate meaningfully. Investment in data infrastructure (e.g. geospatial asset databases, climate-financial risk data hubs) will pay off by reducing the noise and bias in aggregate measures. Essentially, better micro data = more reliable macro aggregates.

- ✓ **Conceptual and normative clarity.** We should recognize what our aggregates represent and what they omit. GDP, for example, is not a welfare measure; adding natural capital accounting is one corrective. Similarly, a "1.5°C warming" target, while a useful aggregate goal, omits information about regional extremes – climate policy should incorporate complementary targets or bounds (perhaps something like "no region experiences >X°C increase" in addition to the global mean target). A future framework might set a vector of climate goals (temperature plus adaptation/resilience metrics) rather than a single number. In short, be clear about values and judgments embedded in aggregates.
- ✓ **Interdisciplinary collaboration.** Economists, climate scientists, and complexity theorists need to continue cross-pollinating methods. For instance, machine learning could be used to approximate the results of complex simulations in a formula that policymakers can use – essentially automating aggregation. Or insights from climate science about fat-tailed damage distributions could inform financial stress test scenarios to include more severe edge cases.

We outlined some frameworks like hybrid modelling (embedding detailed micro models within macro models), iterative feedback aggregation in stress tests, and dashboards of metrics. Validating those through research and practice is a future task. Pilot projects where, say, a central bank and climate scientists jointly model a nation's climate financial risk using both macroeconomic models and detailed ABMs could provide blueprints for others.

Another proposal is developing **scenario ensembles** not just for climate variables but for economic responses – instead of a single "orderly vs disorderly transition" narrative, consider multiple plausible pathways of human response and aggregate outcomes under each. This is conceptually similar to multi-model ensembles in climate science, and helps bracket aggregate uncertainty by scenario diversity.

Adaptive frameworks: Given the uncertainties, policy frameworks need to be adaptive. That means regularly updating aggregated risk assessments as new data come in (for example, incorporate observed climate extremes to adjust damage functions – an updating of aggregated risk). It also means having contingency plans for aggregate outcomes outside the expected range. For example, central banks could say "if system climate losses appear to be trending beyond our X% of capital threshold, we will implement Y measure." This is setting triggers based on aggregated metrics but acknowledging ahead of time the possibility of tail events.

Educational aspect: Decision-makers at all levels should be literate in reading aggregated indicators with a critical eye. This calls for making the presentation of uncertainty and distribution as routine as presenting the mean. The more stakeholders appreciate the nuance, the more appetite there will be to invest in robust measurement and to avoid oversimplified conclusions.

To sum up, *"what gets measured gets managed"* – but only if measured correctly. Misleading aggregates can lead to mismanagement. The literature and practices reviewed here show a clear trend: towards integrated, multi-faceted risk assessment. We see a future where an economic report or climate risk report begins with a rich set of indicators (with uncertainty bands), tells a coherent

story using them, and bases recommendations on this full picture. Achieving that will require continued innovation in methods and, importantly, the will to apply them in policy despite their complexity.

The road ahead for research and policy development includes:

- **Developing better theoretical aggregation theorems** for cases with near decomposability plus known exceptions (to guide when representative models are valid vs when ABMs are needed).
- **Building open-source simulation platforms** that allow users to plug in micro data and obtain aggregate risk distributions, lowering the barrier to sophisticated analysis.
- **Creating forums and standards** for sharing best practices on everything from how to aggregate climate scenarios across models to how to reflect model uncertainty in aggregated outputs (maybe by presenting ranges across models, as we discussed).
- **Encouraging policy exercises** like scenario gaming that explicitly address cross-sector aggregation – e.g., a national climate risk drill where different ministries (energy, agriculture, finance) input their sectoral assessments and a central team aggregates them to identify gaps (like something falling through the cracks at aggregate level).

In the end, tackling issues as sprawling as climate change or ensuring financial stability in a changing world is akin to solving a giant puzzle. Each piece (each dataset, each model, each sector) provides part of the picture. The job of researchers and policymakers is to fit these pieces together without forcing them into the wrong place or leaving gaps. That means sometimes aggregating, sometimes disaggregating, and always questioning whether the picture we see is true to the pieces that form it.

References:

Bank of England. (2021). Climate Biennial Exploratory Scenario: Financial risks from climate change. Bank of England.

Bank of England. (2022). System-wide stress test results. Bank of England.

Bank of England. (2025). Monetary policymaking at the Bank of England in uncertain times. Bank of England.

Boltzmann, L. (1872). Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen. Sitzungsberichte der Akademie der Wissenschaften, 66, 275–370.

Boskin, M. J., Dulberger, E. R., Gordon, R. J., Griliches, Z., & Jorgenson, D. W. (1996). Toward a more accurate measure of the cost of living: Final report to the Senate Finance Committee. U.S. Government Printing Office.

Dasgupta, P. (2021). The Economics of Biodiversity: The Dasgupta Review. HM Treasury.

European Central Bank. (n.d.). Distributional Wealth Accounts. https://www.ecb.europa.eu/pub/economic-research/research-networks/html/research_networks_dna.en.html

Federal Reserve Board. (n.d.). Distributional Financial Accounts. https://www.federalreserve.gov/releases/z1/dataviz/dfa/

Financial Stability Board (FSB). (2021). FSB roadmap for addressing climate-related financial risks. FSB

Fisher, F. M. (1971). Aggregate production functions and the explanation of wages: a simulation experiment. The Review of Economics and Statistics, 305-325.

Fisher, I. (2005). The making of index numbers: A study of their varieties, tests, and reliability. Cosimo Classics. (Original work published 1922)

Fleurbaey, M. (2009). Beyond GDP: The quest for a measure of social welfare. Journal of Economic Literature, 47(4), 1029–1075.

Fritzsche, K., Schneiderbauer, S., Bubeck, P., et al. (2014). The Vulnerability Sourcebook: Concept and guidelines for standardised vulnerability assessments. GIZ (Deutsche Gesellschaft für Internationale Zusammenarbeit).

Fritzsche, K., Schneiderbauer, S., Zebisch, M., et al. (2021). Climate Risk Sourcebook. GIZ (Deutsche Gesellschaft für Internationale Zusammenarbeit).

Gibbs, J. W. (1902). Elementary Principles in Statistical Mechanics. Yale University Press.

Gorman, W. 1953. Community Preference Fields. Econometrica 21: 63–80.

Kirman, A. (1989). The intrinsic limits of modern economic theory: the emperor has no clothes. The Economic Journal, 99(395), 126-139.

Kirman, A. (1992). Whom or what does the representative individual represent? Journal of Economic Perspectives, 6(2), 117–136.

Lamperti, F., Bosetti, V., Roventini, A., & Tavoni, M. (2019). The public costs of climate-induced financial instability. Nature Climate Change, 9(11), 829–833.

MacKay, R. S., & Robinson, J. (2018). Markov chain aggregation and coarse-graining. Entropy, 20(8), 605.

Network for Greening the Financial System (NGFS). (2022). NGFS Climate Scenarios for central banks and supervisors. NGFS.

Paisley, J., & Nelson, M. (2025). A Risk Professional's Guide to Physical Risk Assessments: A GARP Benchmarking Study of 13 Vendors. Global Association of Risk Professionals (GARP) https://www.garp.org/risk-intelligence/sustainability-climate/comparing-climate-risk-251023

Piketty, T., Saez, E., & Zucman, G. (2018). Distributional national accounts: Methods and estimates for the United States. Quarterly Journal of Economics, 133(2), 553–609.

Reiss, J. (2021). Measurement in economics. In The Routledge Handbook of Philosophy of Economics (pp. 123–137). Routledge.

Shaikh, A. (1974). "Laws of production and laws of algebra: the humbug production function." The review of economics and statistics, 115-120.

Simon, H. A. (1962). The architecture of complexity. Proceedings of the American Philosophical Society, 106(6), 467–482.

Sonnenschein, H. (1972). Market excess demand functions. Econometrica, 40(3), 549–563.

Sonnenschein, H. (1973). Do Walras' identity and continuity characterize the class of community excess demand functions? Journal of Economic Theory, 6(4), 345–354.

Sonnenschein, H. (1982). Price adjustment and aggregate excess demand. Econometrica, 50(2), 539–547.

Stern, N. (2006). The Economics of Climate Change: The Stern Review. Cambridge University Press.

Winsberg, E. (2012). Values and uncertainties in the predictions of global climate models. Philosophy of Science, 79(5), 830–841.

Zebisch, M., Schneiderbauer, S., Fritzsche, K., et al. (2021). Climate Risk Sourcebook. GIZ (Deutsche Gesellschaft für Internationale Zusammenarbeit).

Zhu, J. (2025), "Microfoundations in measurement and theory", Deutsche Bundesbank unpublished memo

The Smith School of Enterprise and the Environment (SSEE)

SSEE was established with a benefaction by the Smith family in 2008 to tackle major environmental challenges by bringing public and private enterprise together with the University of Oxford's world-leading teaching and research.

Research at the Smith School shapes business practices, government policy and strategies to achieve net zero emissions and sustainable development. We offer innovative evidence-based solutions to the environmental challenges facing humanity over the coming decades. We apply expertise in economics, finance, business, and law to tackle environmental and social challenges in six areas: water, climate, energy, biodiversity, food, and the circular economy.

SSEE has several significant external research partnerships and Business Fellows, bringing experts from industry, consulting firms, and related enterprises who seek to address major environmental challenges to the University of Oxford. We offer a variety of open enrolment and custom Executive Education programmes that cater to participants from all over the world. We also provide independent research and advice on environmental strategy, corporate governance, public policy, and long-term innovation.

For more information on SSEE please visit: www.smithschool.ox.ac.uk

Oxford Sustainable Finance Group

Oxford Sustainable Finance Group are a world-leading, multi-disciplinary centre for research and teaching in sustainable finance. We are uniquely placed by virtue of our scale, scope, networks, and leadership to understand the key challenges and opportunities in different contexts, and to work with partners to ambitiously shape the future of sustainable finance.

Aligning finance with sustainability to tackle global environmental and social challenges.

Both financial institutions and the broader financial system must manage the risks and capture the opportunities of the transition to global environmental sustainability. The University of Oxford has world leading researchers and research capabilities relevant to understanding these challenges and opportunities.

Established in 2012, the Oxford Sustainable Finance Group is the focal point for these activities.

The Group is multi-disciplinary and works globally across asset classes, finance professions, and with different parts of the financial system. We are the largest such centre globally and are working to be the world's best place for research and teaching on sustainable finance and investment. The Oxford Sustainable Finance Group is part of the Smith School of Enterprise and the Environment at the University of Oxford.

For more information please visit: sustainablefinance.ox.ac.uk/group