Comments on: *Bringing it all together: Science and modelling priorities to support international climate policy.*

This is a timely article, and I welcome its use of the *EGUSphere* to facilitate an open and accessible discussion on important topics and the invitation to comment. If others engage with reasoned opinions, and if the authors respond to these in similarly reasoned ways, the manuscript can make a valuable contribution to the development of scientific priorities.

As it stands, the article is an imperfect assemblage of what appears to be the individual research programmes of its European authors who collectively have considerable experience in a particular sphere of climate modelling, and its relationship to policy This is a useful starting point. However, in many places the paper reads as if it is: (a) a scientific assessment; and/r (b) somehow speaks for a group larger than the authors themselves. These deficiencies are compounded by an insufficiently critical analysis both in regard to the authors' specific past experience, and of important initiatives by large-segments of the community that precede the submission of this manuscript.

On who the article speaks for: The word international is used nearly three dozen times, mostly in the sense of talking about the needs or experiences of either an "international modelling community" or an "international policy community". This results in the impression that the authors are trying to speak for these 'communities'. Even if such communities existed, how can the author's purport to represent them? Likewise, the institutional affiliation combined with the articles normative tone can give the impression that the authors are speaking for the institutions they are affiliated with. As someone who can speak for my institution, I can say that those authors affiliated with the MPI-M are not representing an institutional view, as we were never consulted on these issues.

A nod is given to the Global South (note both words should be capitalized), quite literally as an after thought in the last paragraph. Given the exclusively western European author list many will see this as tokenism. The authors cannot change who they are, but the deficiencies as might occur in the presentation of the research programme from such a limited set of experiences should be more explicitly acknowledged and ways for addressing them suggested. EVE, which was built from a much more diverse (but still far from perfect) group of scientists, technologists, and practitioners, presents ways to address these issues, by developing state of the art infrastructure in under-resourced regions, this and other proposals could be a basis for a specific discussion of the issues.

Simply by clearly framing the article as a commentary developing the ideas and views of its conceptualizing European authors, and acknowledged their shared history (which I presume, in leading and working together on European projects funded with the objective of supporting successive past phases of CMIP), would go a long way to addressing the above issues.

The above brings me, to another critical point. The article emphasizes the importance of coordination, as if this can happen by force of will. Most of the ideas, and challenges it puts forth are real, but could very easily have been written ten, or even, twenty years ago. The idea that we need more coordination with scenarios, more rapid throughput to users, more ensembles, higher resolution, better models, better uncertainty quantification, etc are all good and fine, but not new. Decades of national and European projects, encompassing hundreds of millions of Euro have been conducted under the pretense that they would overcome these challenges. CMIP6 was estimated by the CMIP panel to have cost €3 billion of funding, yet even its most eloquent advocates have struggled to communicate the scientific accomplishments that merited the effort (Lamarque 2022, Meehl 20023). Given this, the question become: What went wrong and why can it now go right? A starting point to answering this question is offered be the many recent published analyses. (see e.g., Stevens (2024), Palmer and Stevens (2024), Jakob et al. (2023), Slingo et al., 2022). Their answer as to what went right and what has changed are as follows:

- Copernicus Climate Change Services
- Destination Earth
- EVE
- Computational capacity enabling qualitatively new approaches to modelling.
- Machine Learning (but not in the sense it is referred to in this article).

Among these, DestinE and resolution are the most substantial, and EVE the most ambitious.

The authors might object that each of the above is mentioned in their article, but the first three only superficially so. There is expertise on the author team to treat these proposals and projects more seriously.

Resolution is substantially addressed, but the conclusions drawn are in dissonance with the evidence presented. Likewise, ML is mentioned many times (more than 20) but still as an undeveloped promise with few concrete results, and even in this realm the cutting edge applications as discussed in a series of review article (e.g., Hoefler et al., Bauer et al., Bauer et al) are overlooked.

I would prefer for the authors to build on their vast experience, in coordinating and contributing to projects like ESM2025, CRESCENDO, CMIP, to critically review what was achieved, what went wrong, and what is different now? Doing this in the context of ongoing activities (DestinE, NextGEMS, EERIE) as mentioned above would strengthen the article and give the readers, and those who might draw different conclusions, a line of reasoning that can be discussed, and hopefully resolved.

A lack of specificity poses another problem, in that the article is unclear as to whether it is arguing to continue the practice of providing a climate service (advice to the international policy community) using a research infrastructure, or if it is arguing for an operational activity as has been proposed in a number of high-level commentaries. This is a fundamental point as the literature claims that the failures (challenges) that the article articulates are a direct consequence of trying to provide an operational service using a research infrastructure (Bauer et al., 2021, Slingo et al., 2022, Jakob et al., 2023 and Stevens et al., 2024 Stevens 2024). Do the authors think this analysis is incorrect? Are they really of the mind that the challenges their manuscript poses can be met by coordinating a research infrastructure? If so it seems reasonable to expect them to explain why this hasn't worked over the past two decades, and three CMIP cycles, and what has now changed that will make it more effective in a fourth attempt. Alternatively if they indeed are proposing an operational service, then more explicitly building upon and extending the ideas presented by Bauer et al., Slingo et al., 2022, Jakob et al., and Stevens et al., 2024, would make this more tangible. For an article that advocates working together, the absence of a critical discussion of open proposals and ongoing activities by their colleagues makes the authors seem uninformed.

The modelling challenge:

In addition to addressing the challenge of processing information, from scenarios to global models, to impacts, and back, the article again emphasizes the need to improve the suite of tools, particularly the modelling, that is central to this information processing. I suspect there is no disagreement on this point. The real question is however, what strategies are likely to be most successful, and why now?

Throughout our careers we have undoubtedly all heard that models need to improve. I have reviewed and participated in scores of projects that are motivated by the need for improved parameterizations. A common approach has been to learn from high (km-scale and finer) resolution. I've organized and led some of the largest modelling exercises and field campaigns dedicated to these activities. As touched on by my recent CMIP essay (Stevens 2024), the unfortunate reality is that this has not produced any real examples of improvements to parameterizations that systematically improve all models, i.e., the same change across a score of models that lead them all to be better.[1] The efforts for sure have greatly deepened our understanding of how low clouds work. Likewise we now appreciate the important ingredients for representing rainfall, and extremes, as well as for a host of other processes. In most cases we understand the failure of fundamental assumptions common to efforts to apply parametric solutions (e.g., lack of scale gap, assumption of equilibrium) so as to preserve the computational economy of coarsely resolved approaches. These findings end up reinforcing the null hypothesis, which is that parameterization is a model specific form of error compensation, so that particular instances of model improvement are likely to be an example of overfitting (how else to explain the Zhu et al (2023) results. This means that even the meager progress across CMIP phases is, where not related to resolution, not robust It means that overfitting is real, which makes me (and I suspect many others) skeptical that machine learning will be more helpful than misleading as a way to better (rather than just more efficiently) represent the effects of sub grid-scale processes.

Hence the question becomes what has changed? The main thing is the ability of increased computational capacity to tap scales that allow for structurally different simulation systems. This article has, buried within it, many insights that reinforce this point, through its highlighting of a small subset of the numerous and specific examples of how storm-and-eddy resolving capabilities improves simulation quality. These examples stand in contrast to unspecific references to the importance of parameterization development, machine learning, or emergent constraints. And while deep in its bowels the article recognizes the importance of resolution for representing impacts (downscaling) and physics (upscaling), it still misses the

---

[1] Across all parameterizations, and apart from the long-tail mixing introduced by Louis in the late 1970s, or perhaps the Gent-McWilliams eddy parameterization from the 90s, it is hard to identify a single parameterization development that led to systematic improvement across models. Even the systematic benefits of GM are disputed.

key role resolution plays in linking to data (observations). More importantly, its message is systematically undermined by repeated references to the need to improve resolution only to the limit that parameterized models can comfortably access, i.e., (10-20 km) scales rather than the breakthrough scales of 5-10 km and finer.

What is the rationale for this? Are too many of the authors unaware of the state of the art? DestinE has already begun to broach scales where we can represent new physics. Multi-decadal to centennial multi-model simulations can be run with full earth system models (carbon cycle, dynamic vegetation, aerosol are already coupled and ice-sheets pose no particular difficulty) at scales ranging from 2.5-5 km globally. The first scenarios are running on Lumi now, at 5 km, globally, with a throughput of 0.33 SYPD and larger, on about 5% of that machine. In the coming year machines twice as large will become available in Europe. This means that dedicating just one of the existing Tier-0 machines would deliver more than 2000 simulated years per year. Wait a year and we will have twice as many resources. But even now the computational capability is enough for small ensembles of multi-model multi-decadal projections (30-50yr) at 5km (horizontal) grid spacings. The computing capacity is there, as are the models, albeit not all models. Do the authors really believe that the problem is not important enough to use these technologies, or that before using them we need to wait until everyone's model can use them?

This doesn't mean that km-scale global models are the only thing that needs attention, but by being more specific as to what this capability means for existing approaches, and how these approaches could be adapted in light of the new capabilities, any claim to support existing approaches sounds self serving. From my point of view regional models will be very important for wide-scaling (i.e., exploring scenarios to enable machine learning interactivity), and models with the much coarser (20km-200km) resolution will continue to be important for: multi-centennial to millennium scale simulations, for interpreting the results from more physical (highly-resolved) models, and perhaps to improve global scale ensemble inflation techniques with machine learning. Unfortunately the article leaves the reader, at least this reader, with the impression that we should rather double down on a hand whose best cards were played long ago, i.e., in CMIP3.

More minor points:

1. the article gives the impression that CMIP is synonymous with Scenario MIP. And sometimes by only mentioning specific MIPs when not referring to the DECK or the scenarios it reinforces the idea that CMIP is about the scenarios, and the 'community' MIPS are an afterthought. Although this thinking is strongly reflected in the way CMIP presently presents itself, this assumption should be made explicit, or avoided.

2. ensembles are often mentioned, but it is not specifically explained why new ones are still needed. Now that we have a reasonable characterization of ensemble variability, what is to be learned by running new ensembles. Does ensemble variability across large-ensembles differ greatly? If so do we understand why, and how new ensembles might help us better identify the correct answer?

3. considerable attention is devoted to uncertainty, but the words ring empty when there is no strategy to estimate it, and there is no effort to separate it from model spread. Confusing uncertainty with model spread and the spread of model errors does the reader, and our science, a disservice.

4. the focus on negative emission scenarios is interesting, but I missed the question as to why we specifically need to run ensembles of ESMs to address this issue. Before embarking on this endeavor one should answer the question of what the emulators and well tuned simple box models miss and, for this process, what an ESM adds, and why. And, for what it adds (including for the trivial answer of unreliable spatial patterns) why should we believe it. This raises the broader question as to why aren't emulators enough for policy?

5. not addressing the really pioneering advances of the Japanese colleagues in understanding the role of resolution for important atmospheric processes is an enormous oversight. A starting point is the Satoh et al., review article, the contributions it cites, and some since them (e.g., Takasuka et al.), merit more attention.

In retrospect, the minor points (particularly points 2, 3 and 4) constitute a further major point, which is that the article risks being seen as championing a culture of repetition, for repetition's sake. If we do things again, there have to be specific and compelling reasons for doing so, to do otherwise is not scientific.

Summary:

The article is timely, taps considerable expertise, but is insufficiently critical and reflective. My specific concerns can be summarized as follows:

1. Address who the authors are, and how the article should be read, i.e., as a common position statement of the undersigned authors. Or are they really representing something broader, and if so what and how?
2. Critically reflect on what, based on the authors' experience, went wrong (or right) in the past, what is now different, and why and how does that lead us to expect we can do better?
3. Draw the consequences from the experience that only proven path to better models is to better resolve the underlying physics. Explain why, given that decades of efforts to address long-standing systematic biases through improved parameterizations have failed — even for cases where we know the answer (e.g., Fiedler et al., 2019) — should they suddenly succeed now?
4. Constructively engage recent proposals (e.g., EVE, Slingo et al, the Scientific Advisory Board of the WMO, the Royal Society Briefs for COP, Jakob et al), which make the case for operationalizing the process of climate information provision for policy and applications.
5. Explain how the scientific community can best benefit and strengthen considerable investments in initiatives such as DestinE and C3S. (In this regard, and on a specific but important point, the Jakob et al. article is cited, but incorrectly. Jakob et al. argue quite clearly for an operational activity, not a quasi-operational repurposing of a research infrastructure.)

Addressing these five points, and framing the presentation in a way that is consonant with this analysis would be more scientifically exemplary. Even as a commentary or perspective the article should strive toward this standard.

Bjorn Stevens,
Max Planck Institute for Meteorology

Selected (own) References:

(Apologies for almost exclusively sharing only own references, but an effort to be more comprehensive is, as I see it, the job of the authors.)

Bauer, P., Stevens, B. & Hazeleger, W. A digital twin of Earth for the green transition. *Nat. Clim. Chang.* **11**, 80–83 (2021).

Bauer, Hoefler, Stevens, Hazeleger, Digital twins of Earth and the computing challenge of human interaction, Nature Computer Sci., (2024) in press

Bauer, P. *et al.* Deep learning and a changing economy in weather and climate prediction. *Nat Rev Earth Environ* **4**, 507–509 (2023).

Fiedler, S. et al. Simulated Tropical Precipitation Assessed across Three Major Phases of the Coupled Model Intercomparison Project (CMIP). Monthly Weather Review 148, 3653–3680 (2020).

Hoefler, T. *et al.* Earth Virtualization Engines: A Technical Perspective. *Comput. Sci. Eng.* **25**, 50–59 (2023).

Jakob, C., Gettelman, A. & Pitman, A. The need to operationalize climate modelling. Nat. Clim. Chang. 13, 1158–1160 (2023).

Lamarque, J.-F. Planning for the next phase(s) of CMIP. (2022). Report to Joint Scientific Committee (JSC) of WCRP.

Meehl, G. A. The Role of the IPCC in Climate Science. in Oxford Research Encyclopedia of Climate Science (Oxford University Press, 2023). doi:10.1093/acrefore/9780190228620.013.933.

Palmer, T. & Stevens, B. The scientific challenge of understanding and estimating climate change. Proceedings of the National Academy of Sciences 116, 24390–24395 (2019).

Slingo, J. et al. Ambitious partnership needed for reliable climate prediction. Nat. Clim. Chang. 12, 499–503 (2022).

Stevens, B. A Perspective on the Future of CMIP. *AGU Advances* **5**, e2023AV001086 (2024).

Stevens, B., et al.,: Earth Virtualization Engines (EVE), Earth Syst. Sci. Data, in press, 2024.

Satoh, M. et al. Global Cloud-Resolving Models. Curr Clim Change Rep 5, 172–184 (2019).

Takasuka, D. et al. How Can We Improve the Seamless Representation of Climatological Statistics and Weather Toward Reliable Global K-Scale Climate Simulations? Journal of Advances in Modeling Earth Systems 16, e2023MS003701 (2024).