The authors are to be commended for condensing what was invariably a great many ideas into a coherent and well structured perspective. Their manuscript brings together a wealth of experience to help explain why they find perturbed parameter ensembles to be useful. I also liked the selection of Figures, which were well constructed and informative. The Authors' views, presented in this manner, will be of general interest to the community and serve as a touchstone for further discussions. I also found no issues with their Perspective's reading of the literature, in which sense I think the article is scientifically sound. I would thus simply offer the authors the benefit of my views, which they can respond to as they see fit, but which need not entail changes to their manuscript.

I'd begin with a few overarching comments:

The first is that the manuscript leaves it to the reader to infer the authors' view of modelling. My inference is that they think it is to provide an instrument to guide policy, i.e., the authoritative view, a crystal ball. But there are other views, and these have an implication for how one interprets the manuscript. It would be helpful if these views were more in the forefront when presenting the authors ideas.

My view of modelling is that it is a tool we create to advance our reasoning. Models aren't for prediction, rather predictions are what we do to test our reasoning and this reasoning is informed by the use of models. This leads to my favored interpretation of the manuscript, as an elaborated way of saying that people should play with their models to understand why they do what they do and use this to assess the reasoning that the models helped develop and which can't be more directly tested.

Another view, is that the point of modelling is to establish authority (reliability measures this). This sounds negative, and it is, but it is also necessary in practice, for instance to develop policy. Hence it is also a valid view, which the manuscript appears to adopt It then proceeds to give the impression that PPEs can be coordinated in a formal way, to better establish the authority (reliability) of a model, rather than its suitability as a tool for developing an argument.

This leads to my second overarching point. Assuming that we are talking about models as tools for prediction, then I don't see how the author's ideas could be implemented. The paper certainly presents many arguments as to why their ideas have merit But it assumes that they could be implemented to improve the practical use of models. I think this assumption is false.

My skepticism has a couple of origins. One, is that the ideas are not new, and such approaches have not worked in the past, at least not as applied to Earth system models. So why would they work now? To address this it could help to consider if past failures occurred for structural reasons, i.e., related to the fact that the models (at least the parameterizations) aren't really physical, but rather structurally different statistical fits, to a developing understanding of the data. Or were past failures for cultural reasons, i.e., we lack the discipline and organizational skill needed to implement the approach. And if one or the other, what do we need to do differently, and why are we capable of doing so.

Part of the difficulty is this idea of matching observations. In model development we generally have an idea of when a simulation is closer to some observable quantity as compared to another simulation, and this often guides both structural and parametric choices. However in Earth system modelling, it almost never happens that a change to a model is closer to all observations. Likewise it almost never happens that the same change across all models gives the same improvement. It would seem that both are required for the author's programme to make sense, and neither is. The better the example of how my thinking is wrong, the more optimistic I would be about the proposed research programme.

To address the above it would be helpful to work out one or two examples of how the author's views could be implemented. The detail and specificity of the examples will be important, as until now all one has are vague references to studies that point in a given direction. Some questions that should be addressed by the examples would be: How do the results depend on what parameters are chosen? Models have thousands of parameters. Is the idea to expose them all and vary them all? Is that possible? Is it worth it? How to deal with the fact that similarly named parameters have different meanings in different models, and that many parameters are hidden? How to determine plausible parameter ranges of unphysical parameters? How many models are needed for this programme to work? How to deal with the fact that structural uncertainty is

grossly undersampled (we all more or less use the same model, e.g., Shaw and Stevens, *Nature*, 2025). And most importantly, what would the result look like at the end?

Some more specific comments:

- In section 3 I would have welcomed a more specific discussion of what was learned, rather than what was done. Learning generalizes.
- Despite the definition of terms, which I very much liked, the authors used the word uncertainty quite loosely. Also, if a parameterization is based on a false assumption, how can the parameters it uses have a correct value, and if they can't then what does parameter uncertainty really mean. This all seems predicated on the idea that the model is structurally correct and we know it is not. Hence I think it is not correct to think of structural and parametric uncertainty, but rather one should speak of structural and parametric *sensitivities*.
- Mauritsen et al., (*J. Adv. Model. Earth Syst.* 2012) was the first study to openly discuss the strategy of tuning a CMIP model, and it encountered quite a bit of resistance from our own community who felt that this was opening our field to disingenuous criticism. The strategy we adopted at that time was very much cognizant of the idea of equifinality. Hourdin et al., (2012) is similar and came out at the same time. The later Hourdin et al., article on the "Art and Science" was an outgrowth of these more foundational contributions (which incidentally was initiated by S. Bony, who recognized the importance of these ideas and, as these things go, wasn't every properly acknowledged.)
- For Fig 4. How does one reject implausible estimates? Somehow this happens magically in the box labeled observations, but it is the crux of the matter. If this were possible then it should also be possible with the MME and that raises two questions: why hasn't it been done in the past (i.e., in IPCC assessments of the CMIP ensemble); and whether the best estimate would end up being different? In other words maybe MMEs adequately sample parameter and structural sensitivities.

Bjorn Stevens

2025-10-04, Berlin