

Reviewer #1

We are grateful for the constructive comments, which have helped to greatly improve the manuscript, together with comments by the 2nd reviewer.

This paper introduces a novel framework for assessing bias in climate models by mapping model state variables directly into the observation space for comparison with in situ data. The key novelty is that, in contrast to traditional climate model evaluation, which often maps observations into model space and compares against archived model output, this approach evaluates model output in a form that is more naturally aligned with the observing system. For kilometer-scale modeling, the authors argue that the traditional approach can obscure fine-scale structure and introduce avoidable errors, which OBSALL is designed to mitigate. Here, they introduce OBSALL v1.0, an online evaluation tool that takes the model state and produces outputs suitable for comparison with specific observational data products.

An inherent limitation of this approach, which the authors acknowledge, is that the modeling team must know in advance which observational datasets they wish to compare against before running the simulations. The authors also note that this is a first version of the system and that the framework can be extended to incorporate additional observing systems. This reflects an important shift in how kilometer-scale climate models must be developed and evaluated, since the cost and complexity of these simulations require model setup and evaluation planning to proceed together. In that sense, the paper highlights a broader change in workflow philosophy, where it is no longer practical to adopt a “store everything and analyze it later” approach at full native resolution.

As a disclaimer, my expertise is not in assessing model bias, so I am not comfortable reviewing the scientific rigor of Section 4 in detail. However, I can comment on the technical details and the broader impact of this work on the community. I think this paper provides a valuable instrument in the toolbox being developed for kilometer-scale modeling. As a community, we will need to revisit many traditional paradigms and determine how well they translate to the era of storm-resolving global models. This paper is well written and well presented. I believe it will benefit the storm-resolving model community and can serve as a resource for other modeling centers considering how to assess bias in their models. The authors have made a clear effort to discuss and address the potential limitations of the new technique, and the paper is explicitly positioned as a first step that may encourage other modeling centers to consider similar tools. I am comfortable recommending publication, provided the authors address the grammar corrections listed below. I also include a few broader questions that do not necessarily need to be addressed for publication, but would likely strengthen the paper.

General Questions

- I may have missed it in the text, but I do not believe the authors discuss how this tool could be adopted outside of the three models they use in the paper. For OBSALL v1.0 to have a broader impact on the climate modeling community, it would be helpful to discuss the process for integrating this tool into a more generic climate model workflow. Have the authors thought about the portability of their technique and code? Is it possible to create a generic implementation that could be applied more universally, similar to many physics parameterizations that have found widespread use across multiple climate models?

We agree that this has not been discussed clearly. Subsection 2.3 is revised to contain a new paragraph at the end of 2.3 (highlighted yellow) to address the point raised. OBSALL v1.0 contains build-in options for off-line and on-line access. These enable off-line access to evaluate stored simulation data and on-line access for run-time monitoring and evaluation of ongoing simulations. The latter would require some technical efforts. The proposition of a “generic solution” is interesting and will fuel our development although it is not further discussed in the revised manuscript.

- A little more discussion would also help on the intuitive argument for why kilometer-scale models really cannot rely on the traditional approach of mapping observations onto the modeled state.

Thanks for this comment. Km-scale requirements are indeed central for our argumentation. As a result (and also because similar request by Reviewer #2), Abstract, Section 1 (Introduction) and Section 5 (Discussion) are re-written to cover this point (and in response to other points raised by the Reviewers).

- There is some discussion of the limitations of the metrics presented in the study, see the paragraph near line 588. It would be helpful to explicitly point out that, although the evidence in the paper suggests observation-based comparisons are better, a more thorough evaluation would require more observational datasets and longer simulations to better capture trends.

Thanks for this comment. It is true that limitations of the observation-space approach are not collectively explained, especially in the context of traditional approaches. Due to this Comment (and in response to comments by Reviewer #2), limitations of both traditional evaluation methods and our approach are now covered in Sections 1 and 5 of the revised manuscript.

Technical comments

There are a few grammar issues the authors may want to address:

- Line 17: “enables to produce” missing a subject, consider “enables the framework to produce”
- Line 18: “a range on impact-sector applications” unclear, perhaps “a range of impact-sector applications”
- Line 25: “is available run-time” should be “is available at run time”
- Line 34: “contain also significant” should be “also contain significant”
- Line 38: “at local level” may need “the”, as in “at the local level”
- Line 38: “informs directly the scale of interest” could be revised to “directly informs the scale of interest”
- Line 57: “At process-level” should likely be “At the process level”
- Line 86: “constitutes as step change” should be “constitutes a step change”
- Line 113: “is rare in climate modeling” may be too strong, since there are established efforts to map model output to observation space, such as COSP. It may be worth clarifying how COSP and similar tools differ from OBSALL
- Line 145: “if it was not for” should be “if it were not for”
- Line 168: “which enables to access” should be “which enables access to”
- Line 242: “reports at least four times a day the basic mete...” should be revised to “report the basic mete... at least four times a day”
- Line 250: “subtracting linear trend” should be “subtracting a linear trend”
- Line 283: “missing observations are also omitted” , it would help to indicate how common this is, for example, whether this affects 1%, 10%, or another fraction of the data
- Line 523: “rising a flag” should be “raising a flag”
- Line 581: “allows to evaluate observation-based...” needs revision, perhaps “allows observation-based evaluation” or “allows us to evaluate...”
- Line 596: “inform on digital” should be “inform digital”

Thanks for these notes, especially the important comment about the too strong expression (“rare”). Introduction of the revised manuscript contains a review of traditional evaluation approaches (model-space and “signal space” methods, including the COSP -type signal simulation tools), their strengths

and weaknesses, the special limitations related to km-scale evaluation, and how OBSALL sits in this landscape and manages to remedy certain long-standing weaknesses. This addition makes the discussion more balanced and respectful regarding earlier work. This revision is in-line with the comments by Reviewer#2, which point in the same direction as this comment by Reviewer #1: improve review of state-of-the-art.

In addition, comment related to number of missing data is now addresses in the text and in a new Appendix.

Figure comments

The figures would be easier to interpret if they were formatted to remain readable in black-and-white printing and for color-blind readers.

- Figure 2: Consider using different line styles for the three models, and possibly different marker types, to improve readability in black and white
- Figures 4, 7, 8, and 12: The same comment applies
- Figure 8: What do the “x” markers represent? They are not identified in the key or the caption and should be explained
- Figure 9: For black-and-white viewing, circles for heatwaves and squares for coldwaves would make the figure easier to parse

We thank Reviewer #1 for these detailed tips/suggestions for improving those Figures that display station-specific data. Figures [2, 4, 7, 8, 10, 11, 12] of the manuscript and [1, 3, 5, 7, 9, 11, 13, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39] of the supplement are revised. We have:

- modified the line styles so that different models have different line styles,
- made the grey colour of the uncertainty representation slightly lighter so that the line styles for different models would be more contrasting,
- used different markers for the observations and models in order to enable also black-and-white viewing

In Figures [8] of the manuscript and [11, 13, 27, 39] of the supplement we also added texts identifying observations and models in the right-hand side bar charts in order to enable black-and-white viewing. In Figures [9] of the manuscript and [15, 16] of the supplement we changed the marker style of coldwaves to squares in order to enable also black-and-white viewing.

The “x” markers in Figure 8 are now explained in the caption (denote observations that were used to fit the Gumbel distribution to extreme events having different return periods in the original observational data).