Hybrid forecasting: using statistics and machine learning to integrate predictions from dynamical models

Response to Reviewer 1

Anonymous Referee #1, 20 Oct 2022

Thank you for the opportunity to review Slater et al. "Hybrid forecasting: using statistics and machine learning to integrate predictions from dynamical models". Overall, I find this to be a timely and informative review. However, I do have a variety of comments, detailed below. I recommend at least a minor revision, if not a major revision.

We are grateful to Reviewer 1 for their positive and helpful comments on our manuscript. Their comments are copy-pasted below verbatim in black font, and our replies are in blue font. We label the comments in the following manner: "R1.C1" indicates Reviewer 1, Comment 1.

R1.C1: My biggest concern in reading this paper is the number of different models and approaches etc. that are discussed. The paper is full of acronyms (so Table 2 is certainly helpful) such that I routinely found myself lost in the details and trying to remember the bigger picture or category that the details were supporting. If I'm someone coming to this review trying to figure out where to start with hybrid modeling, I think I would really struggle. How would I begin? Would I choose a model/paper from Table 1? How would I discriminate or know how to choose among the myriad of options? If the authors can provide some answers or guidance to these types of questions, I think it would be very helpful. Also, if there is any way to more clearly emphasize the main points even among all the details.

We appreciate this opinion, and we agree with the Reviewer that the paper should provide a clear introductory overview of the different types of hybrid models for someone new to the field. We propose to add a new paragraph that more clearly outlines where a user could begin (depending on their aim) and articulates the main characteristics of each hybrid model type. We propose to base this paragraph on an expanded and improved version of the current Table 3 (see response to **R1.C17**).

R1.C2: Terminology is really important in this paper. Can you please provide some definitions of the differences between physics-based vs. conceptual models?

Yes, we agree entirely. We have now included definitions of physics-based, conceptual, and dynamical models upfront in the first paragraph of the revised manuscript.

R1.C3: One question I had was whether any hybrid schemes are currently operational. But, this is partially answered in line 93. Also wanted to see what the authors think it would take to make these models operational, which is partially addressed in the conclusion. Any further details that can be provided on this topic would be greatly appreciate (i.e., are there ANY examples of operational hybrid schemes? And if so, can they serve as pilot projects? i.e., what can we learn from their implementation that might help hybrid schemes become more widely used?).

We appreciate this is an important point and we will add discussion of operational hybrid schemes in the revised manuscript. For instance, the US Climate Prediction Center runs an objective consensus climate forecast that uses ensemble regression to combine multiple dynamical and statistical forecasts into one. The International Research Institute for climate and society (IRI) has a multi-model calibrated prediction based on three SubX models. Lastly, the Google flood forecasting model is also operational. We will include a discussion of what can be learnt from the implementation of these hybrid schemes.

Beyond these operational hybrid examples, there are also cases of hybrid forecasting where the statistical part of the forecast is run separately by a stakeholder using dynamical forecast outputs from the producing centre -- these examples are not always visible as a single 'hybrid' activity but are operational nonetheless. We will include some examples.

Finally, one important point that we will discuss more explicitly in the revised text is that almost all climate-scale hydroclimate projection is hybrid. This is because some kind of statistical processing is almost always applied to an ensemble of CMIP outputs (though the projections may not necessarily be 'operational').

R1.C4: Lines 100 and on list many hybrid models… but not all the references are in Table 1 as well. Any reason? (e.g., Miller et al., 2021)
Initially, we only presented a representative selection of models in Table 1 because we were concerned that it might become a very long table. In the revised manuscript, we will add further key hybrid papers that are discussed in the manuscript and we will justify our selection.

R1.C5: Section 2.4 seems to have a different focus than what is indicated on line 122.
Thank you for spotting this; the text has been revised accordingly ("and hybrid forecasts including a conceptual hydrological model").

R1.C6: The grammar of the sentence spanning lines 122-124 isn't quite correct. Same for the sentence spanning lines 273-274.
Thank you - both sentences have been updated.

R1.C7: Lines 243: seems like a concluding statement (summarizing the overall point of the paragraph) is needed here.
We agree, and a concluding summary statement will be added.

R1.C8: Line 249: the reference to Madadgar et al., 2016 – where was this study applied?
The study was applied to the southwestern United States. The text has been updated to reflect this.

R1.C9: Lines 264-266: Is this sentence a description of "mode-matching"? And if so, can that be made clear. If not, please provide a brief idea of what mode-matching is.
We have included a definition and updated the citations.

R1.C10: Line 409: by "national" does that mean the United States?
There are different CAMELS datasets for different countries, including the United States, United Kingdom, Chile, Brazil, Australia, France, and Switzerland (available soon). We will clarify which countries these datasets are available for, in the revised text.

R1.C11: Line 440: what does "surface water" mean?
The term "surface water" at line 440 referred to a paper by Rözer et al. (2021) on pluvial flood forecasting. We will clarify the text.

R1.C12: Lines 454-461: this paragraph, especially the last sentence, seems to imply there are no limitations to hybrid models.
Thank you - we did not intend to give this impression and will revise the wording. The sentence "*The previous 'limitations' can thus now be seen as challenges*" suggested that they are no longer limitations. We will reframe this point to show that the limitations *"are indeed challenges…"*, and we will ensure the limitations are more clearly highlighted in the revised manuscript. These include, for instance, challenges related to physics-guided ML designs, difficulties in assimilating human influences, or the 'curse(s) of dimensionality' (problems of data sparsity, multicollinearity, multiple testing and overfitting) mentioned by Reviewer 2.

R1.C13: Lines 491-509: are these paragraphs in the correct place? The information presented within seems to go in Section 2.1 on pre- and post-processing.
These two paragraphs have been moved to section 2.1, and will also be shortened a little, as that section is now quite long.

R1.C14: Lines 598-599: this is a really important point that I'm glad was made (i.e., the marginal improvement might be not worth the effort). It seems to me that dealing with this issue is critical to making hybrid schemes more widely accepted. Is there any way we can determine a priori the marginal improvement (without having to build both models in parallel and then compare)? For example, the Mai et al. (2022) study in line 616 – would be good to comment if the demonstrated superiority was enough to justify the extra effort.
Yes, we agree that this is an important but tricky point, as it is hard to know *a priori* how much added value can be obtained without first building a hybrid model and benchmarking the results. We agree that the Mai et al. (2022) study is the first of its kind in providing such a detailed intercomparison of modelling approaches, and it suggests the effort related to using ML is justifiable. Although the findings do help sway the field, they are for

simulation, rather than forecasting (for which there are more steps needed). In other words, it is still a jump to speculate that ML provides improvements for prediction.
Another important point is that the development of successful hybrid methods for short-term forecasting does not automatically imply success for medium range, subseasonal-to-seasonal, or decadal forecasting. In the revised text, we will include a discussion of various issues associated with the implementation of operational hybrid schemes, such as the shift of expertise, and the potential shift of computing architecture when implementing GPU-requiring ML techniques. We will more clearly discuss the fact that a lot of hybrid operational forecasting is currently being implemented, as described in the response to **R1.C3**. In many cases, the dynamical producing centre draws a line before the 'tailoring' statistical part, which the stakeholders implement.

R1.C15: Table 1: (a) Are any of these operational? (b) Any rationale for inclusion/exclusion of studies in this table? (c) Can you add another column that describes how the statistical and dynamial models are combined? (d) Regarding column headings, in the text, "data-driven" seems to be the most generic term (lines 25-26) but here the column header is "statistical" model (and elsewhere, "empirical" is used). Again, the importance of terminology in this paper. (e) Would this table become slightly easier to digest if it was first sorted by predictand type (i.e., streamflow vs. reservoir, etc) and then horizon? I'm not sure, but I think that predictand is a larger category (and what I would first be interested in), then horizon.
(a) We will include a number of operational examples, such as those mentioned in our reply to **R1.C3**. We are revising the text to reflect that hybrid hydroclimate forecasting is a form of operational practice already.
(b) We sought to cover different types of dynamical and statistical models, different ranges, and different variables. Including all the available papers would be too many, but we will ensure that we have a representative sample of all the different study types (and will attempt to provide a classification of the types that exist).
(c) This is an excellent suggestion. Depending on space, we will either add a column in Table 2, or provide further examples in Table 3 (which we plan to develop; see response to **R1.C17**).
(d) The column heading has been updated to "data-driven".
(e) We have made the choice to sort by horizon, but will assess how the table looks when sorted by predictand then horizon, and make a decision. Thank you for the useful suggestions!

R1.C16: Some acronyms that are not defined anywhere: RCP8.5, FV3GFS (this is just the name of the atmospheric model?), PREVAH (also a model name?)
The definitions of these acronyms have all been added to Table 2: Representative Concentration Pathway 8.5 (high-emissions warming scenario); Finite-Volume Cubed-Sphere Global Forecast System (global atmospheric model); and Precipitation-Runoff-Evapo-transpiration Hydrotope Model. We will also check the text to make sure we have not accidentally missed any other acronyms.

R1.C17: Table 3: (a) Shouldn't "coupled" be included here also, since it is discussed in the text. (b) I find it interesting that Lee et al. (2002) is a primary reference for two of the options (serial and parallel) – given that it is now 20 years ago. Is that because it was such a foundational paper? Either way, can a more recent reference also be provided? As a corollary comment: It would be nice to have a discussion in the text of when these approaches were first tried (what was the foundational paper) on hydroclimate variables.
We propose to update Table 3 to provide a more comprehensive overview of the different types of hybrid structure that exist and will use this revised table to better clarify and describe the approaches in the main text.
(a) We will update Table 3 to include the 'coupled' approach, and we will make sure the terms 'coupled' and 'parallel' are more clearly defined (for instance, we will distinguish the term 'coupled' from coupled dynamical models and clarify that one-way post-processing from dynamical to statistical is not coupling).
(b) Lee et al. (2002) happened to use and compare both those terms, but we agree that more recent references should be added in here, and will update the table accordingly.
We will also add and discuss foundational papers, such as Glahn and Lowry (1972) on postprocessing using model output statistics (MOS). One challenge is that different terminology is used in different fields, and the approach is not always referred to as "hybrid forecasting", so it can be difficult to find older papers.

Glahn, H. R., & Lowry, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting. Journal of Applied Meteorology and Climatology, 11(8), 1203-1211.

R1.C18: Figure 1: A few comments/questions on this graphic: (a) Please explain if the coloration of the boxes has any meaning. (b) Aren't large-scale predictors etc. also inputs to the hybrid forecasting scheme (not just dynamical predictors) – in other words, the straightforward left-to-right is not actually quite so straightforward? (c) Bottom middle: shouldn't it be "hydroclimate model" rather than "hydrological model" to be more general? Thank you for helping us make the figure more intuitive.

(a) The colour of the boxes indicates the broad type of prediction scheme and serves to help the reader see how the top two schemes (rows) are combined in the third scheme (bottom row, reflecting hybrid prediction); we have clarified the figure caption accordingly (please see revised caption below).

(b) Yes, large-scale predictors can also be used as inputs, but would likely be issued from dynamical predictions or dynamical reanalyses (e.g., using large scale principal components to identify predictors) in the case of a hybrid forecast (although some observations might be employed too). We will update the figure caption to make it clear that multiple different types of combinations are possible.

(c) Yes, we agree that the bottom middle box would be better with the term "hydroclimate model" and have updated it accordingly; thank you for spotting this.

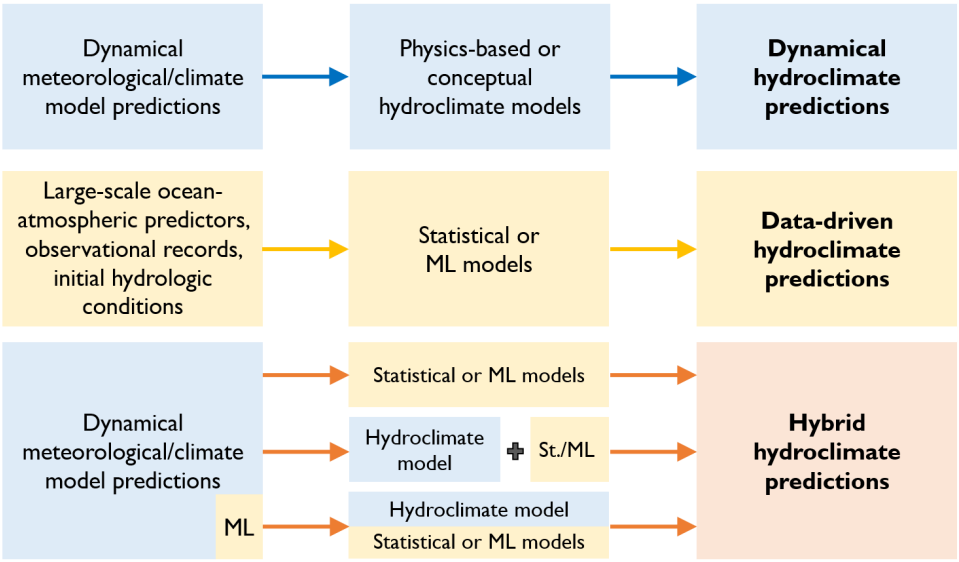

Figure 1. Defining hybrid hydroclimate forecasting and prediction. "Hydroclimate" refers to a range of variables defined in the text (including streamflow). The top row indicates traditional dynamical hydroclimate predictions (blue); middle row is data-driven predictions (yellow) and bottom row is hybrid predictions (red, with three examples of hybrid structure from top to bottom: statistical-dynamical, serial, and parallel, with an example of ML-based post-processing of dynamical model output in the bottom left box). The figure provides a simple example, but more complex schemes are possible, including e.g., a mix of observations and predictions in the left column.

R1.C19: Figure 2: So, you obtain one value each for JJA, then take the max? Could be clarified in the caption text. The maximum summer discharge is the largest of the 92 daily values in the June-July-August period. The caption has been revised to state this explicitly.

Thank you for this constructive review!