

Advancing Subseasonal Forecasting with Machine Learning

Corresponding Author: Dr Judah Cohen

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)

Summary:

This manuscript describes an approach to probabilistic bias correction of subseasonal (ensemble) forecasts and demonstrates how this approach can improve state-of-the-art traditional and AI-based forecasts. The underlying statistical concepts borrow ideas from existing bias correction methods, but the methodology proposed here impresses with its generality and strong performance within the setup of ECMWF's 2025 real-time forecasting competition. The manuscript is well written, and the included figures are clear and mostly support the claims made in the article.

General comment:

The only aspect where in my opinion the generality of this approach is not fully demonstrated is related to the fact that it is strongly centered around the correction of (cumulative) quintile probabilities. There is nothing wrong with this in itself, but the authors should discuss how the resulting forecast CDFs could be used to predict the probabilities of truly extreme events. In Section 2.5, extreme events are defined as "those verifying in the lowest or highest quintiles". By definition, such an "extreme" then occurs in 2 out of 5 weeks, and while I understand the challenges of both forecasting and verifying truly extreme events at subseasonal forecast lead times, I find it a bit of an overstatement to associate the demonstrated accuracy gains with extreme weather with "the highest societal impact" and to claim that they "have the potential to profoundly benefit society" (l. 174-175). To better support such a statement, I feel that the authors should

- explain how probabilities for exceeding high thresholds - both fixed and expressed as high(!) climatological quantiles - can be derived from the rather coarsely discretized CDFs that the proposed methodology produces
- verify and compare the performance within a truly high impact event setup

Specific comments:

- Figures 2 and 5 suggest that the skill reported for the subseasonal precipitation forecasts is primarily driven by a large RPSS over regions that I would expect to be extremely dry. I find this a bit surprising because in my experience it is especially hard to beat a well-constructed precipitation climatology in areas with little to no precipitation. Can you add some more discussion of why PBC works especially well in that situation and what the associated performance gains mean from an impact perspective?

- Can you generally provide some more discussion - perhaps backed up by further analyses - of which component of the PBC methodology explains its strong performance compared to alternatives (PoET and the methods mentioned in references 34-38)? Since neither the discretization into climatological quantiles nor the blending of forecast with lagged observations and/or climatology are completely novel ideas (e.g. Scheuerer et al., 2020, Mouatadid et al., 2023), it would be interesting to get a better understanding of which of these concepts primarily explain the success of PBC

- l. 362-365: Can you explain how these variable-specific weights were determined and why the additional PoET bias

correction is useful for mean sea level pressure but not useful for precipitation?

References:

Mouatadid, S., Orenstein, P., Flaspohler, G., Cohen, J., Oprescu, M., Fraenkel, E., Mackey, L.: Adaptive bias correction for improved subseasonal forecasting. *Nature Communications* 14(1) (2023)

Scheuerer, M., Switanek, M.B., Worsnop, R.P., Hamill, T.M.: Using Artificial Neural Networks for Generating Probabilistic Subseasonal Precipitation Forecasts over California. *Monthly Weather Review* 148 (2020)

(Remarks on code availability)

The code base for this project is quite comprehensive and I haven't reviewed everything in detail, but I can confirm that this is a very well organized and well documented GitHub repository with a detailed description of how the method can be used and how the results of the paper can be reproduced.

Referee #2

(Remarks on code availability)

I have looked over the code, and read the README page. I have not attempted to install and run the code, but the necessary instructions seem to be there.

(Remarks to the Author)

This manuscript presents a probabilistic bias-correction framework (PBC) that improves the probabilistic skill of ECMWF subseasonal forecasts and AI-based forecasting systems. The paper is well written, technically sound, and demonstrates consistent improvements across multiple variables, lead times, and forecasting systems. The strong performance in the AI Weather Quest competition further highlights the operational relevance of the approach. I support publication after revision.

My main concern is the framing of the results. In several places, the manuscript appears to overstate both the predictability of the atmosphere and the implications of the reported skill gains. The atmosphere remains a chaotic system with a well-established intrinsic predictability limit. Statements such as "deterministic forecasting systems can now provide accurate point estimates of the weather up to two weeks in advance" (lines 43-46) are too strong and should be revised. Similarly, the statement that systematic errors are "predictable and correctable using machine learning" (lines 61-62) could be interpreted as implying a larger increase in predictability than is actually demonstrated.

The results are significant and operationally relevant, but the absolute skill improvements are generally modest. Throughout the manuscript, the authors emphasize relative gains such as "doubling" or "tripling" forecast skill. While these statements are correct, I recommend consistently reporting both absolute and relative improvements, particularly in the abstract and discussion, to provide a more balanced assessment of the practical significance of the results.

I would also encourage the authors to provide a clearer breakdown of the contribution of the different components of PBC. The framework combines Debias++, Persistence++, isotonic projection, and a final ensemble step, but only the performance of the complete system is reported. Quantifying the contribution of each component would help readers understand why the method works and which elements are primarily responsible for the reported improvements.

Finally, while the operational results are compelling, PBC is fundamentally a post-processing framework. The manuscript would benefit from a clearer discussion of what is novel relative to existing bias-correction and calibration approaches and why the demonstrated improvements represent an important advance for subseasonal forecasting.

Overall, this is a strong and potentially impactful paper. My comments primarily concern the interpretation and presentation of the results rather than the methodology itself.

Specific comments:

- Line 25 - 26: Forecast skill degrades not only because of model error accumulation but also because memory of the atmospheric initial conditions is progressively lost.
- Lines 43 - 46: "Accurate point estimates of the weather up to two weeks in advance" is too strong and should be revised.
- Line 49: Replace "uncertainty about the future" with "forecast uncertainty."
- Line 52: Replace "high skill" with "forecast skill."
- Lines 61-64: Rephrase to avoid implying that machine learning can overcome intrinsic predictability limits; the results demonstrate improved calibration and reduced systematic error.

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)

My comments on the previous version of this manuscript have been addressed, in particular, forecasts for more extreme events are now generated and evaluated and the contributions of the different components of the proposed methodology is now evaluated and discussed. Previously unclear statements and figures have been clarified.

(Remarks on code availability)

The methodology hasn't changed significantly compared to the original version, so I have not re-evaluated the code.

Referee #2

(Remarks to the Author)

The authors have addressed all my concerns. I think the manuscript should be published as is.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response to Reviewers

We are encouraged by the supportive comments regarding our manuscript, and we thank the reviewers for their detailed and constructive feedback. In our revision, we have attempted to address all suggestions and concerns, and we provide point-by-point responses below.

Reviewer 1

Reviewer Point P 1.1 — The only aspect where in my opinion the generality of this approach is not fully demonstrated is related to the fact that it is strongly centered around the correction of (cumulative) quintile probabilities. There is nothing wrong with this in itself, but the authors should discuss how the resulting forecast CDFs could be used to predict the probabilities of truly extreme events. In Section 2.5, extreme events are defined as "those verifying in the lowest or highest quintiles". By definition, such an "extreme" then occurs in 2 out of 5 weeks, and while I understand the challenges of both forecasting and verifying truly extreme events at subseasonal forecast lead times, I find it a bit of an overstatement to associate the demonstrated accuracy gains with extreme weather with "the highest societal impact" and to claim that they "have the potential to profoundly benefit society" (l. 174-175). To better support such a statement, I feel that the authors should

- explain how probabilities for exceeding high thresholds - both fixed and expressed as high(!) climatological quantiles - can be derived from the rather coarsely discretized CDFs that the proposed methodology produces

- verify and compare the performance within a truly high impact event setup

Reply: Thank you for this valuable feedback. We have addressed each of these points as follows:

1. In Lines 102–106 we clarify that, while our study focuses on quintile forecasting to align with the format of the AI Weather Quest, the PBC framework can be applied to any level of CDF discretization. As a practical demonstration in Section 2.5, we now follow prior work [1] and define extreme highs as events verifying above the 95th percentile and extreme lows as events verifying below the 5th percentile. We then train PBC to target semidecile (5% quantile) increments rather than quintiles, separately evaluate its skill at forecasting extreme highs and its skill at forecasting extreme lows for each variable in week 3, and compare that performance with raw and debiased ECMWF in Figures 5a and 5b.
2. In Supplementary Appendix A.6 we carry out an analogous analysis for extreme high forecasting and extreme low forecasting in week 4.
3. In addition, in Section 2.5, we now verify and compare the performance of PBC-ECMWF within a truly high impact event setup. In our new Figure 5c, we first evaluate flood forecasting skill across all flood events catalogued by the Global Disaster Alert and Coordination System (GDACS) [2] in the years 2016–2026. For each event, flood forecasting skill is now quantified by the BSS for the top precipitation semidecile in the event bounding box. We find that PBC consistently boosts raw flood forecast skill, while debiased ECMWF routinely degrades it. GDACS additionally categorizes each flood based on its humanitarian impact and labels a flood severe (Orange or Red) if it results in more than 100 deaths or 80,000 displacements [2]. For these high-impact events, operational

debiasing only degrades or matches the BSS of raw ECMWF, while PBC significantly enhances its severe flood forecasting skill.

4. We have also added as Figure S9 an example from the AI Weather Quest in which MicroDuet's real-time week 4 forecast for the lowest temperature quintile anticipated the December 2025 cold air outbreak in the Eastern United States, an extreme cold event missed by debiased ECMWF.
5. We have removed the statement, "Extreme weather has the highest societal impact, and such improvements in forecasting extremes have the potential to profoundly benefit society [3]."

Reviewer Point P 1.2 — - Figures 2 and 5 suggest that the skill reported for the subseasonal precipitation forecasts is primarily driven by a large RPSS over regions that I would expect to be extremely dry. I find this a bit surprising because in my experience it is especially hard to beat a well-constructed precipitation climatology in areas with little to no precipitation. Can you add some more discussion of why PBC works especially well in that situation and what the associated performance gains mean from an impact perspective?

Reply: We apologize for the confusion regarding this point. As we note in Section 4.1, arid grid cells with no climatological precipitation are excluded from our temporal skill evaluations; the same exclusion is carried out in the AI Weather Quest, as any observation for such a grid cell must fall in the highest quantile bin, trivializing the prediction problem. As a result, these arid grid cells do not contribute to the temporal skill measures reported in the paper and hence are not driving our subseasonal precipitation skill gains. In the revision, we have also removed all arid points from our spatial precipitation skill plots to avoid further confusion. In addition, we have added a discussion of the value and impact of improved precipitation prediction (including in semi-arid regions) to Section 3.

Reviewer Point P 1.3 — Can you generally provide some more discussion - perhaps backed up by further analyses - of which component of the PBC methodology explains its strong performance compared to alternatives (PoET and the methods mentioned in references 34-38)? Since neither the discretization into climatological quantiles nor the blending of forecast with lagged observations and/or climatology are completely novel ideas (e.g. Scheuerer et al., 2020, Mouatadid et al., 2023), it would be interesting to get a better understanding of which of these concepts primarily explain the success of PBC.

Reply: Thank you for this suggestion! We have added a new figure (Figure S8) that highlights the individual contributions of the PBC-PoET components and a new discussion (in Section 2.3) of the components of the PBC framework that explain its strong performance compared to alternatives.

Reviewer Point P 1.4 — 1. 362-365: Can you explain how these variable-specific weights were determined and why the additional PoET bias correction is useful for mean sea level pressure but not useful for precipitation?

Reply: Thank you for this recommendation! We now explain in Section 4.3.4 that, to form the MicroDuet model used in the AI Weather Quest, we considered three combinations of the PBC-ECMWF and PBC-PoET forecasts—an equal-weighted average, full weight on PBC-ECMWF, and full weight on PBC-PoET—and selected the combination that maximized 2024 RPSS for each variable and lead time. For precipitation, the skill of PBC-ECMWF was so much higher than that of PBC-PoET that the

ensemble underperformed PBC-ECMWF in 2024. Similarly, for sea level pressure, the skill of PBC-PoET was so much higher than that of PBC-ECMWF that the ensemble underperformed PBC-PoET in 2024.

Reviewer 2

Reviewer Point P 2.1 — The paper is well written, technically sound, and demonstrates consistent improvements across multiple variables, lead times, and forecasting systems. The strong performance in the AI Weather Quest competition further highlights the operational relevance of the approach. I support publication after revision.

Reply: Thank you!

Reviewer Point P 2.2 — My main concern is the framing of the results. In several places, the manuscript appears to overstate both the predictability of the atmosphere and the implications of the reported skill gains. The atmosphere remains a chaotic system with a well-established intrinsic predictability limit. Statements such as "deterministic forecasting systems can now provide accurate point estimates of the weather up to two weeks in advance" (lines 43-46) are too strong and should be revised. Similarly, the statement that systematic errors are "predictable and correctable using machine learning" (lines 61-62) could be interpreted as implying a larger increase in predictability than is actually demonstrated.

Reply: Thank you for this feedback; we have modified the referenced statements in the introduction.

Reviewer Point P 2.3 — The results are significant and operationally relevant, but the absolute skill improvements are generally modest. Throughout the manuscript, the authors emphasize relative gains such as "doubling" or "tripling" forecast skill. While these statements are correct, I recommend consistently reporting both absolute and relative improvements, particularly in the abstract and discussion, to provide a more balanced assessment of the practical significance of the results.

Reply: Thank you for this suggestion! We now reference the modest baseline skill of AIFS in the abstract and report absolute skill numbers alongside statements of relative improvement throughout the main body.

Reviewer Point P 2.4 — I would also encourage the authors to provide a clearer breakdown of the contribution of the different components of PBC. The framework combines Debias++, Persistence++, isotonic projection, and a final ensemble step, but only the performance of the complete system is reported. Quantifying the contribution of each component would help readers understand why the method works and which elements are primarily responsible for the reported improvements.

Reply: Thank you for this recommendation! We now provide a clear breakdown of the contribution of the each PBC component in Section 2.2 and Figure S7.

Reviewer Point P 2.5 — Finally, while the operational results are compelling, PBC is fundamentally a post-processing framework. The manuscript would benefit from a clearer discussion of what

is novel relative to existing bias-correction and calibration approaches and why the demonstrated improvements represent an important advance for subseasonal forecasting.

Reply: Thank you for these recommendations! We now include an expanded discussion of the novelty of PBC relative to existing approaches at the end of Section 2.3 and an expanded discussion of why PBC represents an important advance for subseasonal forecasting in Section 3.

Reviewer Point P 2.6 — Specific comments:

- Line 25 - 26: Forecast skill degrades not only because of model error accumulation but also because memory of the atmospheric initial conditions is progressively lost.
- Lines 43 - 46: "Accurate point estimates of the weather up to two weeks in advance" is too strong and should be revised.
- Line 49: Replace "uncertainty about the future" with "forecast uncertainty."
- Line 52: Replace "high skill" with "forecast skill."
- Lines 61-64: Rephrase to avoid implying that machine learning can overcome intrinsic predictability limits; the results demonstrate improved calibration and reduced systematic error.

Reply: We have revised each of these statements following your recommendations.

Editorial Suggestions

Editorial Suggestion E 0.1 — Technically, the main issue is raised by referee 1: the focus on quintiles, while useful and valid, is incomplete. We support the referee's suggestion and would like to see a quantification of skill for truly extreme events.

Reply: Following the recommendations of Referee 1, we have added a quantification of skill both for truly extreme and truly high-impact events; please see [P 1.1](#).

Editorial Suggestion E 0.2 — Conceptually, the main issue is raised by both referees: the framing and statements of advance require moderation. Doing so seems reasonable and would not affect our editorial interest.

Reply: We have revised our framing and statements of advance following the recommendations in referee points [P 1.1](#), [P 2.2](#), [P 2.3](#), and [P 2.6](#).

References

- [1] N. S. Diffenbaugh, J. S. Pal, R. J. Trapp, and F. Giorgi. Fine-scale processes regulate the response of extreme events to global climate change. *Proceedings of the National Academy of Sciences*, 102(44):15774–15778, 2005. doi: 10.1073/pnas.0506042102. URL <https://www.pnas.org/doi/abs/10.1073/pnas.0506042102>.
- [2] D. Masante, D. Barantiev, E. Destro, S. Paris, C. Proietti, V. Salvitti, M. Santini, and M. Mastrotrunzio. *Multi-hazard early warning system Global Disaster Alert and Coordination System (GDACS) – Methods for anticipation and rapid impact estimation user manual (version 2)*. Publications Office of the European Union, Luxembourg, 2025. doi: doi/10.2760/1461943.
- [3] J. G. Shrader, S. Thies, L. Bakkensen, M. Linsenmeier, and D. Lemoine. Weather forecasts become more important for reducing mortality as the climate warms. *Proceedings of the National Academy of Sciences*, 123(16):e2523372123, 2026. doi: 10.1073/pnas.2523372123.