Review of "From climate projections to trajectory-conditioned local predictions" by Michel et al.
Reviewer: Vincent Verjans

This study develops a machine learning model for predicting decadal mean temperatures accounting for both forced climate change and internal variability. The model uses temporal feature extraction and a U-Net architecture (TC-UNet), and performs monthly-resolved decadal mean predictions for the periods 2030-2039 through 2090-2099 under the SSP2-4.5 scenario. TC-UNet is trained on CMIP6 simulations, and evaluated in a cross-validation (CV) framework excluding model families, where families are determined according to model development centres.

The central focus is on further developing constraining methods for future climate projections. The motivation to account for both the forced response and internal variability is valuable. The approach shows promise as a novel method for assessing the contribution of initial conditions in the near- and long-term climate evolution. The study places a strong emphasis on the predicted uncertainty ranges, and how they compare to the unconstrained CMIP6 ensemble range. However, the method itself lacks a comprehensive representation of uncertainty, and the uncertainty analysis relies exclusively on empirical estimates of out-of-sample error. Another issue is that the current diagnosis of long-range predictability from initial conditions ignores errors in the forced response estimate. Finally, another limitation that should be resolved is the absence of appropriate benchmarks, which makes it difficult to assess the predictive performance of TC-UNet. These issues are detailed as the three Major Comments of my review. I also raise Minor Comments regarding the structure of the manuscript, the methods, and the interpretation of results. I do not provide line-by-line comments at this stage, as I believe that the main concerns should first be addressed. Overall, the current work shows promise, but I believe that methodological concerns need to be addressed before the manuscript can be considered for publication in *Earth System Dynamics*. Nevertheless, I encourage the authors to address these comments and those from other reviewers, as this study could make a valuable contribution. Line numbers in this review refer to those in the pdf version of the manuscript.

**Major Comments**

**1. Treatment of predictive uncertainty**

The uncertainty ranges provided in all Figures and analyses of the manuscript rely on the CV-based error distribution. The central issue with this approach is that the distribution of out-of-sample predictive performance is generally not equivalent to the prediction uncertainty distribution for a new sample. In mathematical notation:

$p(\hat{Y}_k - Y_k | X_k, D^{(k)}_{train}) \not\equiv p(Y^* | X^*, D)$, (Eq. I)

where $k$ denotes a CV batch, the * superscript denotes a new data sample, $X$ denotes the predictors, $D$ the full CMIP6 data set, and $Y$, $\hat{Y}$, $D^{(k)}_{train}$ are defined as in the manuscript. There are several limitations from using $p(\hat{Y}_k - Y_k | X_k, D^{(k)}_{train})$ as an estimate of prediction uncertainty.

(a) The predictive coverage of the CV-based error intervals has not been assessed. Since these intervals have been constructed from the CV errors, by design, they achieve the specified nominal coverage on these errors. However, this does not establish that they achieve the same nominal coverage on independent prediction errors. See for example Fig. 3 of Ribes et al. (2021) and Fig. S4 of Liang et al. (2020) for out-of-sample evaluations of predictive coverage.

(b) As clearly shown in Eq. (I), the CV-based intervals do not depend on the new input sample $X^*$. Consequently, the uncertainty estimates do not account for differences in the characteristics of the target samples. For example, the prediction uncertainty would be identical if the input sample is (i) a simulation from any of the 40 ESMs, (ii) the ERA5 record, (iii) a blended SST-SAT record, or (iv) a paleoclimate reconstruction from the early-Holocene. Although this example is deliberately extreme, I believe that it illustrates the limitation of interpreting the CV-based error distribution as a general measure of predictive uncertainty.

(c) The CV-based error distribution depends on the choice of the CV design. The authors have carefully designed their LOMFO-CV framework. Still, the error distribution would have been different if, for example, the authors had decided to impose more restrictive family distributions by grouping models sharing a same ocean model physical core. Another example would be if CMIP5 simulations had been included. Consequently, the estimated predictive uncertainty ranges would depend on these choices, even if the predictive model and the new input sample remained unchanged. This contrasts with uncertainty frameworks that explicitly account for inter-model covariance and that condition on the new input sample $X^*$, thereby incorporating information about the structure of the model ensemble and the characteristics of the target sample.

(d) The CV-based error distribution does not account for observational uncertainty. While this source of uncertainty is zero in ESM simulations, it does arise when applying TC-UNet to real-world reanalysis data (Sect. 3.4).

(e) The CV-based error distribution does not capture parametric uncertainty. Although the authors average predictions from five TC-UNet instances, they only use the resulting mean prediction. Parametric uncertainty is not propagated, either in the CV-based errors or in the predictive uncertainty.

(f) As clearly shown in Eq. (I), the CV-based error distribution is obtained from TC-UNet instances trained on restricted subsets of the CMIP6 ensemble, $D^{(k)}_{train}$. In contrast, uncertainty ranges provided in the analyses are for TC-UNet trained on the full ensemble, $D$. The respective error distributions may therefore differ. Although the magnitude of this effect is likely minor, it further highlights the distinction between the two terms of Eq. (I).

Taken together, these issues raise fundamental concerns about the interpretation of predictive uncertainty in this study. This is a major concern because a substantial part of the analysis focuses on reductions in future temperature projection uncertainty and comparisons with existing methods for constraining projections. This concern includes the comparison with the results from Ribes et al. (2021) and the framework of Li et al. (2025) in Sect. 3.4.1. In these studies, the uncertainty components from measurement errors, climate model uncertainty, internal variability, input prediction sample, and indistinguishability assumption are propagated into real-world predictions in a mathematically principled manner. The validity of comparing these results with the CV-based error distribution requires further justification.

**2. Predictability arising from initial conditions**

Section 3.2 assesses the extent to which predictive skill stems from the internal variability imprinted in the initial conditions, rather than from the forced response. To this end, the authors compare prediction skill when TC-UNet uses as input the correct versus an incorrect recent trajectory, taken from different ensemble members of the same CMIP6 model. In my opinion, this approach to estimating the predictability stemming from the initial conditions is unnecessarily convoluted.

Evaluating the skill when input and verification are taken from different realisations faces several limitations. First, differences in TC-UNet predictions cannot be uniquely attributed to internal variability, as they may also reflect differences in the estimated forced response. This cannot be assessed since both components are not separable in TC-UNet predictions. The assumption that differences between the predictions compared illustrate differences in predicted "internal variability" is therefore not established (e.g., L304-305 and L320). Second, the evaluation does not compare the TC-UNet skill against the true forced response. Consequently, the comparison is also affected by errors in the TC-UNet predictions of the forced response, which may further differ between the correct and incorrect input predictions.

I do not understand the rationale for this approach, given that the true forced can be approximated by directly using the ESM-specific ensemble mean. Comparing the skill of TC-UNet predictions against that of the corresponding ESM-specific ensemble mean would provide a more direct assessment of the additional predictive skill associated with using the member-specific recent trajectory to constrain internal variability. In my view, this approach is both more straightforward, and better aligned with the initial-condtition predictability that the authors seek to quantify.

**3. The baselines**

The baselines against which TC-UNet performance is compared are introduced in Section 3.1. First, the description of the baselines is incomplete: the 7 baseline methods are described in less than 5 lines of text. This description does not allow the reader to understand how they work or to assess their suitability for the prediction task. I recommend providing a more explicit description of the different baseline methods in the Supplementary Information.

More importantly, there are obvious shortcomings with the baselines. Figure 3 shows that the best-performing baseline is the ensemble mean (EM). The other baseline methods implemented are very weak benchmarks, as they fail to match the performance of the most simple ensemble mean baseline. This may lead to an overly optimistic assessment of the TC-UNet skill.

I recommend that the authors implement two simple baselines based on linear regression. The first is a simple recalibration of the ensemble mean, corresponding to Homogeneous Gaussian Regression as defined by O'Reilly et al. (2020). The second is a multiple regression on the time series of greenhouse gas, aerosol, and natural forcings (e.g., Fig. 3 in DelSole and Tippett, 2026). Both regressions should be performed individually at each grid cell, using the longer CMIP6 series (1850-2025) to adequately constrain the regression coefficients. I believe that these two simple methods would provide more informative benchmarks for evaluating the prediction skill of TC-UNet. Of course, I would also welcome alternative methods proposed by the authors that address the current limitations. My concern here is that comparisons against methods that perform worse than the uncalibrated ensemble mean do not provide convincing evidence of the added value of TC-UNet.

**Minor comments**

**1. Some skill lies in the forced response**

The authors often suggest that skill of TC-UNet stems from conditioning on internal variability (e.g., L265, L601). However, TC-UNet is also able to distinguish between ESMs with different climate sensitivities based on the 46 years of input. A large part of the skill may stem from this ability to predict the forced response trajectory, which the uncalibrated EM baseline does not capture. I recommend that the authors mention more explicitly that this ability to infer the climate sensitivity from the recent trajectory contributes to TC-UNet predictive performance.

**2. SST pattern effect**
It has been demonstrated that the biases of CMIP6 models in reproducing the observed recent warming patterns can lead to biased relationships between warming rates and climate sensitivity (Armour et al., 2024). This SST pattern effect can be accounted for in constraining methods (e.g., Liang et al., 2024). I recommend that this SST pattern effect and its potential implications for TC-UNet predictions be discussed in the Discussion.

**3. Real-world skill**
Estimates of skill from imperfect model predictions do not necessarily reflect real-world predictive skill, even when accounting for climate model families. This limitation has been demonstrated for decadal predictions by Verjans et al. (2026). Systematic model errors may underlie this discrepancy. I recommend that this limitation and its implications for the TC-UNet real-world predictions be discussed in the Discussion.

**4. Expected envelope of fractional coverage (Fig. 9)**
In the analysis of the expected area fraction of predictions falling outside of the empirical uncertainty range, there may be some assumptions not entirely valid (Fig. 9 and L457-462). In particular, the empirical uncertainty range $Q$ (Eq. 7) and the empirical area fraction of local forecasts falling into $Q$ are derived from the same set of CV errors. Consequently, the resulting area fraction does not provide an independent assessment of the expected area fraction for new predictions. This issue may be particularly problematic if the CV-based error distribution is heavy-tailed, as a small number of errors can strongly influence $Q_{5:95\%}$ and, in turn, the % area falling out of $Q_{5:95\%}$. The expected % area envelope is likely sensitive to this assumption of being inferred from the same set of errors. Presumably, the envelope values are optimistic. Indeed, the % area of errors from a new set of predictions falling outside of $Q_{5:95\%}$ is likely to be higher than that of errors from the same set. Thus, the dashed pink lines in Figure 9 should be lower. In this case, the GC5-HH % areas would be beyond the expected empirically-derived envelope. This would indicate an underconfident uncertainty range in predicting the GC5-HH realisation.

**5. Comparison with Ribes et al. (2021)**
In addition to the reservations mentioned in Major Comment 1, I am unclear about how the results of Ribes et al. (2021) are used in this study. First, Ribes et al. (2021) use GSAT in their study, whereas here the authors state that their estimates refer to SAT over land and SST over the oceans (L502-504). Second, I could not find the reported end-of-century warming estimate of 2.68°C (L502) in the original study. I recommend that the authors provide more details on their comparisons with published values, either in the Methods or the Data availability sections.

**6. Forcing TC-UNet with BEST**
TC-UNet is trained using SAT globally, whereas, as described by the authors, BEST combines SST and SAT. The intrinsic characteristics of both temperature fields are therefore different and not interchangeable. This raises the question of whether forcing TC-UNet with BEST provides a meaningful forecast product. As an extreme example, it is clear that forcing TC-UNet with wind speed would lead to meaningless forecasts, as it is a different variable. The mismatch between SAT and combined SST-SAT is less drastic, but still raises questions about the consistency of the training and application data. If the goal is to evaluate spread arising from the choice of input reanalysis product, would it not be more appropriate to force TC-UNet with other SAT reanalysis products (e.g., JRA-55, MERRA-2)?

**7. Some editorial comments**
The manuscript is often unnecessarily verbose and contains repeated material. For example: L62-64 repeats L43, L130-131 repeats the preceding sentence, L288-294 repeats L278-286, L480-484 should belong to a Discussion section, L486-492 repeats Sect. 2.1.2, etc. I encourage the authors to streamline the manuscript by removing redundant material and improving the organisation of information both wihtin and between sections. Also, I encourage to ensure scientific precision in the wording used. For example, the terms prior and posterior are used in ways that are not always clearly defined, which is particularly confusing since they have a well-established meaning in the Bayesian formalism.

**8. A few additional details on TC-UNet**
There are some aspects of the model architecture and training procedure that require clarification.

- What is the loss function used for training?

- The input period is 46 years according to L193 (552 months), but is 31 years according to L707. Which input length is used, and why do these descriptions differ?

- "For each SSP–decade combination, a separate instance of the network is trained. Inputs are 31-year sequences of monthly temperature anomalies, and targets are 10-year monthly climatologies for the corresponding future decade" (L707-708). This sentence is confusing, and it is unclear how the input period is defined for each target decade. For example, does the TC-UNet model for decade 2090-2099 use the previous 31 years (or 46, see above) or does it use a fixed historical period ending in 2025? If it is the former, then why is the 2090-2099 decade labeled as long-range prediction (e.g., Fig. 5), and how is this TC-UNet applied to the observed record, which stops in 2025?

- "degraded held-out MAE across a representative subset of folds" (L210-211). Was this evaluated on the validation or test data?

**9. ESMs included in the estimation of $Q$**
"For each fold $k$ and decade $d$ we draw one error $e_m^{(d)}(s)$ uniformly from each contributing ESM, giving a balanced 40-member sample (i.e., one member for each of the ESMs)" (L166-167). This statement contradicts the description of the LOMFO-CV framework. If, for each fold $k$, the corresponding model family is held out, the held-out family should not contribute to the error distribution for that fold. Consequently, the number of ESMs eligible for the error distribution should be smaller than 40, with the reduction depending on the number of ESMs belonging to family $k$.

# References

K. C. Armour, C. Proistosescu, Y. Dong, L. C. Hahn, E. Blanchard-Wrigglesworth, A. G. Pauling, R. C. Jnglin Wills, T. Andrews, M. F. Stuecker, S. Po-Chedley, I. Mitevski, P. M. Forster, and J. M. Gregory. Sea-surface temperature pattern effects have slowed global warming and biased warming-based constraints on climate sensitivity. *Proceedings of the National Academy of Sciences*, 121(12), 2024. doi: 10.1073/pnas.2312093121. URL `https://doi.org/10.1073/pnas.2312093121`.

T. DelSole and M. K. Tippett. Comparing climate time series – part 6: Testing equality of autoregressive parameters without assuming equality of noise variances. *Advances in Statistical*

*Climatology, Meteorology and Oceanography*, 12(1):73–86, 2026. doi: 10.5194/ascmo-12-73-2026. URL https://doi.org/10.5194/ascmo-12-73-2026.

T. Li, F. W. Zwiers, and X. Zhang. Should we think of observationally constrained multidecade climate projections as predictions? *Science Advances*, 11(20), May 2025. ISSN 2375-2548. doi: 10.1126/sciadv.adt6485. URL http://dx.doi.org/10.1126/sciadv.adt6485.

Y. Liang, N. P. Gillett, and A. H. Monahan. Climate model projections of 21st century global warming constrained using the observed warming trend. *Geophysical Research Letters*, 47(12), 2020. ISSN 1944-8007. doi: 10.1029/2019gl086757. URL http://dx.doi.org/10.1029/2019GL086757.

Y. Liang, N. P. Gillett, and A. H. Monahan. Accounting for pacific climate variability increases projected global warming. *Nature Climate Change*, 14(6):608–614, 2024. doi: 10.1038/s41558-024-02017-y. URL https://doi.org/10.1038/s41558-024-02017-y.

C. H. O'Reilly, D. J. Befort, and A. Weisheimer. Calibrating large-ensemble european climate projections using observational data. *Earth System Dynamics*, 11(4):1033–1049, 2020. doi: 10.5194/esd-11-1033-2020. URL https://doi.org/10.5194/esd-11-1033-2020.

A. Ribes, S. Qasmi, and N. P. Gillett. Making climate projections conditional on historical observations. *Science Advances*, 7(4), Jan. 2021. ISSN 2375-2548. doi: 10.1126/sciadv.abc0671. URL http://dx.doi.org/10.1126/sciadv.abc0671.

V. Verjans, M. G. Donat, C. Delgado-Torres, and T. DelSole. Large potential of performance-based model weighting to improve decadal climate forecast skill. *npj Climate and Atmospheric Science*, 9(1), Apr. 2026. ISSN 2397-3722. doi: 10.1038/s41612-026-01397-6. URL http://dx.doi.org/10.1038/s41612-026-01397-6.