Response to Anonymous Referee #1 comments for manuscript

Summary

The paper proposes and tests a method for downscaling sub-seasonal weather forecasts to improve accuracy and spatial resolution. The approach uses a neural network architecture (3D U-Net) that has previously been used for similar tasks. The forecasts are from the ECWMF ensemble forecast system and the high-resolution data are from PRISM. The method is applied in the Western US. The effect of different input data (ensemble members versus mean, different variable sets) are examined. The authors find that the neural network improves temperature predictions relative to the original NWP forecasts. Results for precipitation are also presented but the quality is more mixed.

We greatly appreciate the reviewer's constructive comments and editorial suggestions, which have considerably improved this manuscript.

High Level Feedback

The topic itself is interesting, and the results shown (particularly for temperature) seem promising. However, it is difficult to evaluate the method due to the omission of important details. Some of the missing methodological details can be deduced by reading the code, but they should be provided in the paper itself.

Thank you for your helpful comment. We have added additional U-Net details to the main text to improve clarity.
Line 108: The 3D U-Net model was trained for 100 epochs using the adam-optimizer with an initial learning rate of 1e-4 and a batch size of 11, selected based on GPU memory limitations (Kingma and Ba, 2017). The network architecture consists of three encoding and decoding blocks, each composed of 3D convolutional layers with 3×3×3 kernels. Average pooling was used for downsampling in the encoder, and transposed convolution was used for upsampling in the decoder. The GeLU activation function was applied after each convolutional layer. To prevent overfitting, we applied early stopping based on validation loss with a patience of 10 epochs. The loss function combines mean squared error (MSE) and spatial pattern correlation, with equal weighting assigned to both components. We chose this combination because each metric emphasizes a different aspect of prediction performance. MSE evaluates the model's ability to reproduce the absolute magnitude of values, while spatial pattern correlation captures the fidelity of the overall spatial distribution, which is particularly important in sub-seasonal forecasting. All configurations were selected through trial-and-error experiments to ensure training stability and generalization capability. These details have been incorporated into the main manuscript for transparency.

In additional to technical omissions, a main concern is the lack of clarity around the purpose and impact of the analysis. The main question, as I understand it, is: what is the effect of using (a) different sets of predictor variables and (b) different ensemble components/aggregations on prediction accuracy? The introduction states that Höhlein et al. (2024) examined (b) and reached approximately the same conclusion as this study. How is this study different? Question (a) seems relevant. However, there is very little discussion of the specific predictor variables (I think they are only mentioned in the SI) and how/why specific variables may contribute to better or worse predictions. Again, it is not clearly explained how

this study differs from the cited Horat and Lerch (2024) or Weyn et al. (2021). Most of the framing of the results and conclusions boil down to "neural network downscaling improves prediction accuracy relative to NWP," which, based on the introduction, seems to already be well-established.

We appreciate this concern and now explicitly clarify our novelty in both scope and approach. Prior studies demonstrated diminishing returns from additional predictors or member-wise inputs in short-range ($\leq$ 120 h) post-processing (Höhlein et al., 2024; Rasp and Lerch, 2018; Schulz and Lerch, 2022). What was unknown is whether these findings extend to sub-seasonal (up to 32 days) daily targets, where regime transitions (*e.g.*, MJO) and aggregation effects alter skill decay. Our 3D U-Net treats lead time as a learnable dimension, enabling transfer of shorter-lead skill to longer leads at daily resolution over complex U.S. terrain.

In addition, we have reframed the impact statement to go beyond "NN improves NWP" by emphasizing operational feasibility and application value. The ensemble-mean, target-only configuration reduces input channels from 400 to 1–2, lowering memory and latency demands by over two orders of magnitude. This enables daily, high-resolution S2S post-processing on commodity GPUs—making the approach viable for routine updates in water, fire, and agricultural decision-support systems. We also highlight robustness to season and land cover type, and candidly note where skill still degrades (*e.g.*, heavy precipitation in complex terrain), which we link to the probabilistic roadmap in the following response.

We have added a concise comparison table to the Introduction contrasting our inputs, horizon, architecture, and findings with cited works, and framed our hypotheses explicitly:

H1: 3D U-Net with lead-time convolutions yields daily S2S skill gains over raw NWP.

H2: At S2S leads, ensemble-mean is equivalent to all-members for deterministic targets.

H3: Auxiliary predictors beyond the target add minimal incremental skill.

This framing and table now appear in the revised manuscript to sharpen the "what's new" message.

Line 42: Prior evaluations of predictor sets and ensemble usage have largely been limited to short lead times ($\leq$5 days) and single valid times (Rasp and Lerch, 2018; Schulz and Lerch, 2022; Höhlein et al., 2024), probing predictability at an instant (Table 1). In contrast, we target sub-seasonal forecasting by supplying sequences of forecast lead times to encode, thereby extending previous findings to lead times longer.

This study enhances predictability in the Western United States through the 3D U-Net-based post-processing that encodes temporal information via forecast lead times and downscaling forecasts to higher spatial resolutions.

Line 283: Furthermore, our results suggest that incorporating additional model-derived predictors or individual ensemble members yields limited improvement in sub-seasonal forecast postprocessing. Notably, the ensemble mean alone performs comparably to using the full set of ensemble components, pointing to a more computationally efficient alternative. These findings extend prior conclusions drawn from short-range forecasting studies (Rasp and Lerch, 2018; Schulz and Lerch, 2022; Höhlein et al., 2024) into the sub-seasonal prediction regime.

Line 306: To move beyond "artificial neural network improves NWP," we emphasize operational feasibility and application value: an ensemble-mean, target-only configuration reduces input channels from 400 to 1–2, lowering memory and latency by more than two orders of magnitude and enabling daily, high-resolution S2S post-processing on commodity GPUs for

routine water, fire, and agricultural decision-support. The approach is robust across seasons and land-cover types, yet skill still degrades for heavy precipitation in complex terrain, addressing these extremes and optimizing the complexity skill balance are priorities. To meet these challenges, we propose advancing into the probabilistic domain.

Table 1: Comparison of the proposed method with previous post-processing studies.

Study	Type	Input		Output	
		Ensemble	Variable	Lead time	Model by lead time
Our study	post-processing	individual member	used additional variables	0 – 32 days, daily	One model for forecast period
Rasp and Lerch, 2018	post-processing	mean, std	used additional variables	48h	Lead time specific model
Schulz and Lerch, 2022	post-processing	mean, std, individual member	used additional variables	0–21 h, hourly	Lead time specific model
Höhlein et al., 2024	post-processing	individual member	used additional variables	wind gust: 6h, 12h, 18h temperature: 24h, 72h, 120h	Lead time specific model
Horat and Lerch, 2024	post-processing	mean	used additional variables	temperature: 3–4W, 5–6 W mean precipitation: 3–4W, 5–6 W accumulate	Lead time specific model

The motivation for the ensemble-based predictors is also confusing. The purpose of an ensemble prediction system is to represent uncertainty, which is not discussed. Also the ensemble members are simulations that, by construction, do not start from the "optimal" estimate of the initial conditions. So it is not surprising that E01 performs worse (unless by "first" ensemble member you mean the control). Interpreting the relative performance of E50 versus E50M requires methodological details that are not provided. But again it is not surprising that the performance is similar given that E50 output is being reduced to a deterministic prediction. It seems like the value of downscaling based on an ensemble would be more in representing forecast uncertainty than improving deterministic downscaled predictions

We agree that the primary added value of an ensemble lies in uncertainty representation. In our deterministic framework, member-wise inputs did not outperform the ensemble mean, suggesting limited exploitation of ensemble spread. We have clarified this in the Discussion and added a short "Future Work" paragraph noting how this could be addressed:

1. Training probabilistic versions of the 3D U-Net (*e.g.*, quantile regression with pinball loss, CRPS-optimized Gaussian output layers; Hersbach 2000; Gneiting & Raftery 2007).

2. Evaluating reliability and spread–skill using Brier scores, rank histograms, and calibration methods (isotonic regression).

3. Testing Bayesian or ensemble-based deep learning methods to better utilize ensemble spread.

While outside the current scope, these extensions are feasible and would align the model more closely with the ensemble's intended purpose. Also, we have also added more detailed methodological descriptions in the manuscript to clarify how the ensemble configurations were used.
Line 191: Our current approach produces deterministic forecasts and therefore cannot fully represent the uncertainty that NWP ensembles are designed to capture. To address this limitation, future work could consider several directions. One option is to train the 3D U-Net to generate probabilistic forecasts, for example via quantile regression with a pinball loss or by predicting parametric distributions (*e.g.*, Gaussian) optimized with the Continuous Ranked Probability Score (CRPS) (Hersbach, 2000; Gneiting and Raftery, 2007). Another is to evaluate the reliability and the relationship between spread and skill using Brier scores, rank

histograms, and calibration methods such as isotonic regression. Finally, more advanced avenues could include modifying the network to produce its own ensemble or adopting Bayesian deep learning frameworks. Line 105: For example, the E50 V8 configuration has 400 input channels, while the E50M V2 configuration has 2 input channels. In the model, the input dimensions are referred to as height, width, and depth, corresponding to lead time, latitude, and longitude, with sizes of 32, 72, and 72, respectively.

Specific Feedback

1. Either the E_{pre} formula or the subsequent description of it is incorrect. You say $E_{pre} = 0$ for perfect predictions, which would require the term inside the square brackets to equal 1. However, it is $2^2 = 4$ when $sigma_{obs} = sigma_{pre}$ and $r_0 = r_i = 1$. It would also be helpful, if possible, to provide some intuition for the terms in this statistic. E.g., why is the standard deviation term squared and the correlation terms raised to the fourth?

   Thank you for pointing that out. There was an error in the original formula, which has now been corrected. This metric was proposed by Yang et al. (2013), and we have followed their method in our study. According to their explanation, certain observational datasets exhibit high spatial pattern reliability, and therefore the correlation term was raised to the fourth power to give it greater weight in the evaluation.

$$E_{pre} = \frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{\left(\frac{\sigma_{obs,i}}{\sigma_{pre,i}} + \frac{\sigma_{pre,i}}{\sigma_{obs,i}}\right)^2 (1+r_0)^4}{4\,(1+r_i)^4}\right] \tag{1}$$

2. Regarding skip connections, for a given level (or spatial resolution) in the u-net, shouldn't there be twice the number of channels in the first layer on the right side of the U as on the last layer on the left (due to the concatenation of feature maps)? This is what is shown in both the Horat and Ronneberger papers. Also, I couldn't determine where the skip connections were implemented in the code but maybe I just missed it.

3. Related to (2), there is no description of the convolution operations (e.g., kernel size). The pooling operation is also not in the body of the text, only in Fig 1 (but not defined). These are scientifically important details considering they control the spatial scales at which information can be extracted.

4. The activation function(s) is also not stated.

5. The loss function is mentioned but not explicitly stated. How are relative weights of the correlation and MSE terms set? Also, point/cell-wise MSE and correlation are closely related so what is the value of including both terms?

   Thank you for the detailed and thoughtful comments. Because comments 2–5 are closely related, we address them together here. As the reviewer correctly noted, the number of channels in the decoder should double at each stage due to concatenation with the corresponding skip-connection features. Our experiments followed this structure. The original schematic may have been confusing because it showed only the post-convolution

channel counts at each layer rather than the doubling that occurs immediately after concatenation. We have revised Figure 1 to clarify the architecture: subscripts beneath each tensor block denote the channel counts, and the operations at each step are indicated by arrows within a dashed box with brief descriptions. We have also added the requested model details to the main text to improve clarity and reproducibility.

The loss function combines MSE and spatial pattern correlation, with equal weighting assigned to both components. The loss function combines mean squared error (MSE) and spatial pattern correlation, with equal weighting assigned to both components. These two metrics capture different aspects of model performance: pattern correlation evaluates how well the model reproduces the spatial characteristics of a region, while MSE measures the absolute error in magnitude. We therefore include both terms in the loss, and this explanation has also been added to the main text.

Line 108: The 3D U-Net model was trained for 100 epochs using the adam-optimizer with an initial learning rate of 1e-4 and a batch size of 11, selected based on GPU memory limitations (Kingma and Ba, 2017). The network architecture consists of three encoding and decoding blocks, each composed of 3D convolutional layers with 3×3×3 kernels. Average pooling was used for downsampling in the encoder, and transposed convolution was used for upsampling in the decoder. The GeLU activation function was applied after each convolutional layer. To prevent overfitting, we applied early stopping based on validation loss with a patience of 10 epochs. The loss function combines mean squared error (MSE) and spatial pattern correlation, with equal weighting assigned to both components. We chose this combination because each metric emphasizes a different aspect of prediction performance. MSE evaluates the model's ability to reproduce the absolute magnitude of values, while spatial pattern correlation captures the fidelity of the overall spatial distribution, which is particularly important in sub-seasonal forecasting. All configurations were selected through trial-and-error experiments to ensure training stability and generalization capability. These details have been incorporated into the main manuscript for transparency.

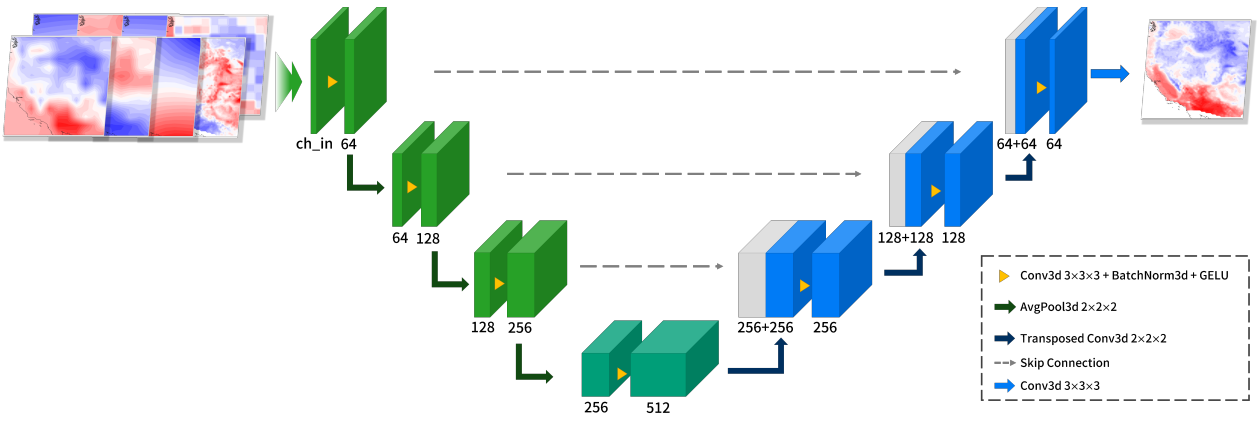

Figure R1: (Figure 1) Schematic of the 3D U-Net architecture adapted for weather forecast post-processing. The model consists of a contracting path (left, green), an expanding path (right, blue), and a bottleneck layer (center, teal), with skip connections (dashed gray arrows) preserving spatial information. Operations between layers are described in the dashed box on the right corner.

6. The claim that pattern correlation and RMSE "quantify the model's ability to capture the spatial patterns" is not justified. Grid cell-wise RMSE is invariant under spatial permutation. The formula or weighting scheme for pattern correlation is not given. From the code it looks like weights are assigned inversely to latitude?

   Thank you for pointing this out. We agree that RMSE is not sensitive to spatial structure, as it is computed pointwise and is invariant under spatial permutation. In contrast, pattern correlation captures the agreement in spatial distribution between the prediction and the observation, and is more sensitive to spatial structure. To clarify, we have revised the manuscript to explicitly distinguish the roles of the two metrics. Also, upon rechecking our implementation, we confirm that pattern correlation is computed using cosine latitude weighting (i.e., proportional to the area of each grid cell), not the inverse of latitude.
   Line 147: Pattern correlation evaluates the model's ability to reproduce the spatial distribution of temperature and precipitation fields, while RMSE quantifies the average magnitude of forecast errors at each grid point.

7. The accuracy assessment would benefit from disaggregation into bias versus "random" errors. Using unbiased RMSE is a way of doing this (see e.g., Entekhabi et al., 2010). In figures 4 and 5, it looks like the downscaled predictions have meaningful biases for certain combinations of variable sets and time steps. For example, panel g4 in Fig. 4 and panel g1 in Fig. 5. If it turns out that the predictions are not "on average" biased, the disaggregation might be less important. However, the presence/absence of bias overall should be mentioned given the visually apparent biases in the figures. This analysis would help identify how much of model performance is coming from downscaling versus simply bias correcting the ECMWF forecasts.

   Thank you for your constructive comment. As shown in the Fig. R2, we calculated RMSE separately for bias and unbiased RMSE. While the NWP forecasts exhibited systematic biases, the deep learning models also showed biases to some extent. In particular, for precipitation, the deep learning models demonstrated a persistent underestimation bias compared to NWP, which we briefly mentioned in Line 164. Nevertheless, we found that the deep learning models still reduced the unbiased RMSE relative to NWP, indicating an improvement in predictive skill beyond bias correction alone. We have revised the manuscript to clarify this point.
   Line 219: However, a consistent underestimation of precipitation is observed across all lead times, with larger biases than those of the NWP model, particularly in coastal and mountainous regions, regardless of the number of input variables. Similar reductions in precipitation during downscaling and U-Net-based post-processing have also been reported in other regions (Xin et al.,2024).

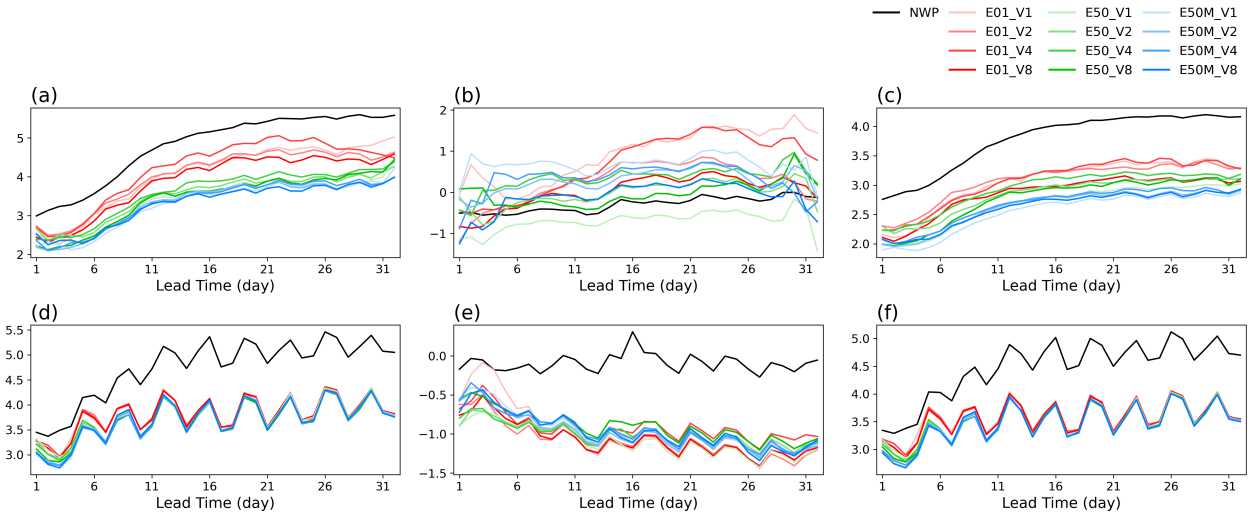

Figure R2: Scores comparing NWP and 3D U-Net models (E01, E50, E50M) for temperature (top row) and precipitation (bottom row) forecasts over 32 days. Columns show (a, d) RMSE, (b, e) bias, and (c, f) unbiased RMSE, respectively

8. The possibility of systematic errors (e.g., season-dependent bias) should be acknowledged/discussed. Even if the predictions are not "on average" biased, there may still be systematic errors. Given that the test period spans one full year, seasonal biases could cancel out such that the predictions appear unbiased in aggregate. The NN predictions may tend toward the overall mean of the training data leading to seasonal biases. I.e., the NN could be introducing systematic biases not present in the NWP output. This may or may not be happing but seems important to consider and rule out.

   Thank you for your feedback. We presented our results using forecast data for 104 initialization dates in 2023. We analyzed the evaluation indices for DJF, MAM, JJA, and SON based on the initialization dates and present these results in Figs R3 and R4. The results show improvements in all performance indices, except for precipitation E_{pre}, across all seasons. For precipitation E_{pre}, performance decreased in spring and summer but increased in winter and spring. Based on these results, we conclude that seasonal bias is minimal. We have added these figures to the supplementary materials and included a discussion of these findings in the main text.
   Line 167: Before conducting a detailed analysis of the results, we examined the potential for seasonal bias and the performance by land cover type. Our findings show improvements in all seasonal evaluation metrics for both temperature and precipitation, except for precipitation E_{pre} in spring and summer (Figs.S4 and S5). This suggests that the enhanced performance is not simply due to the model converging toward the seasonal mean across all seasons. Rather, the improvements reflect the model's ability to capture relevant patterns within each season.

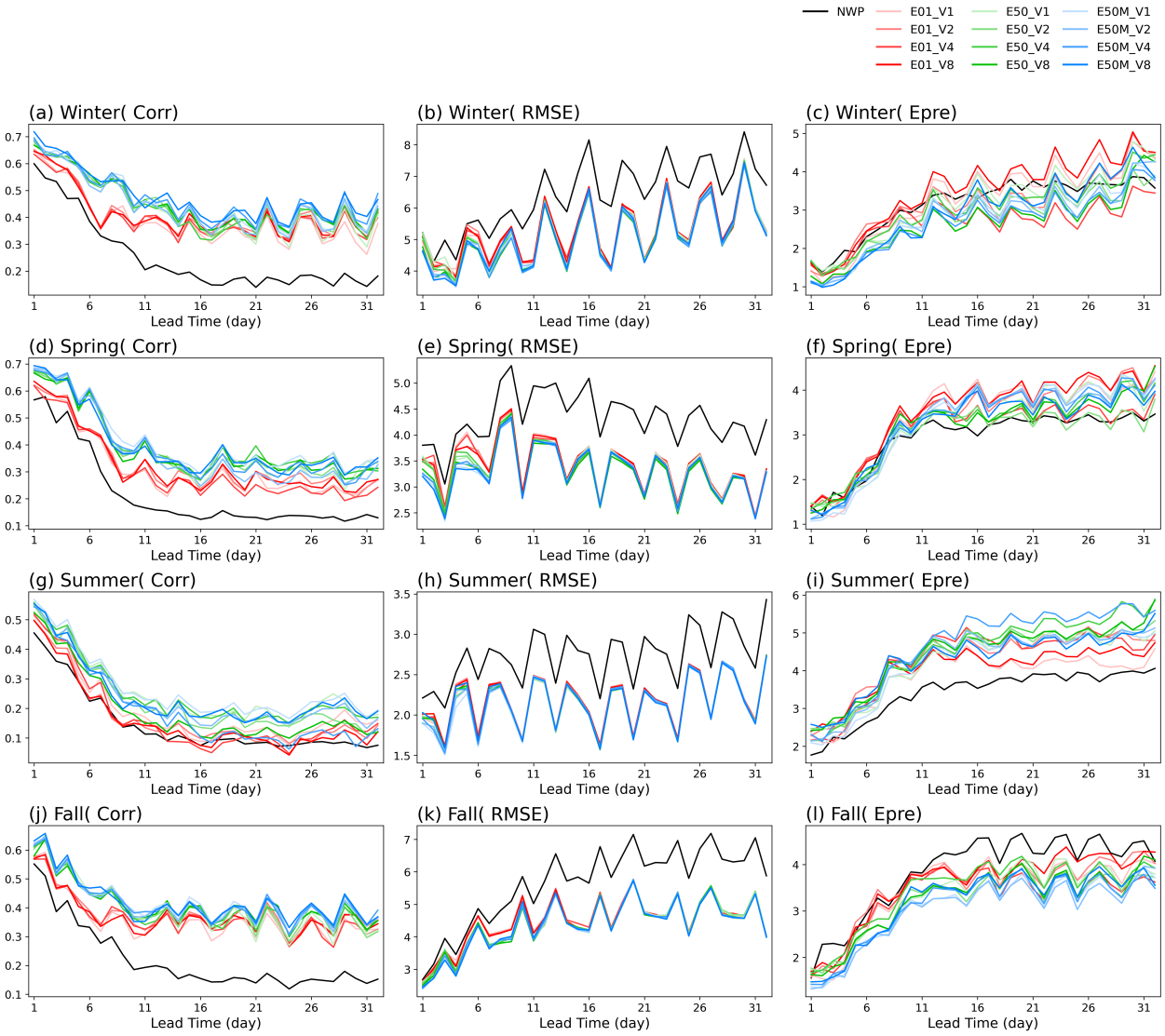

Figure R3: (Figure S4) Seasonal benchmark scores for the Western U.S., comparing NWP and 3D U-Net models for precipitation forecasts. The scores are categorized by initialization dates for DJF (first row), MAM (second row), JJA (third row), and SON (fourth row).

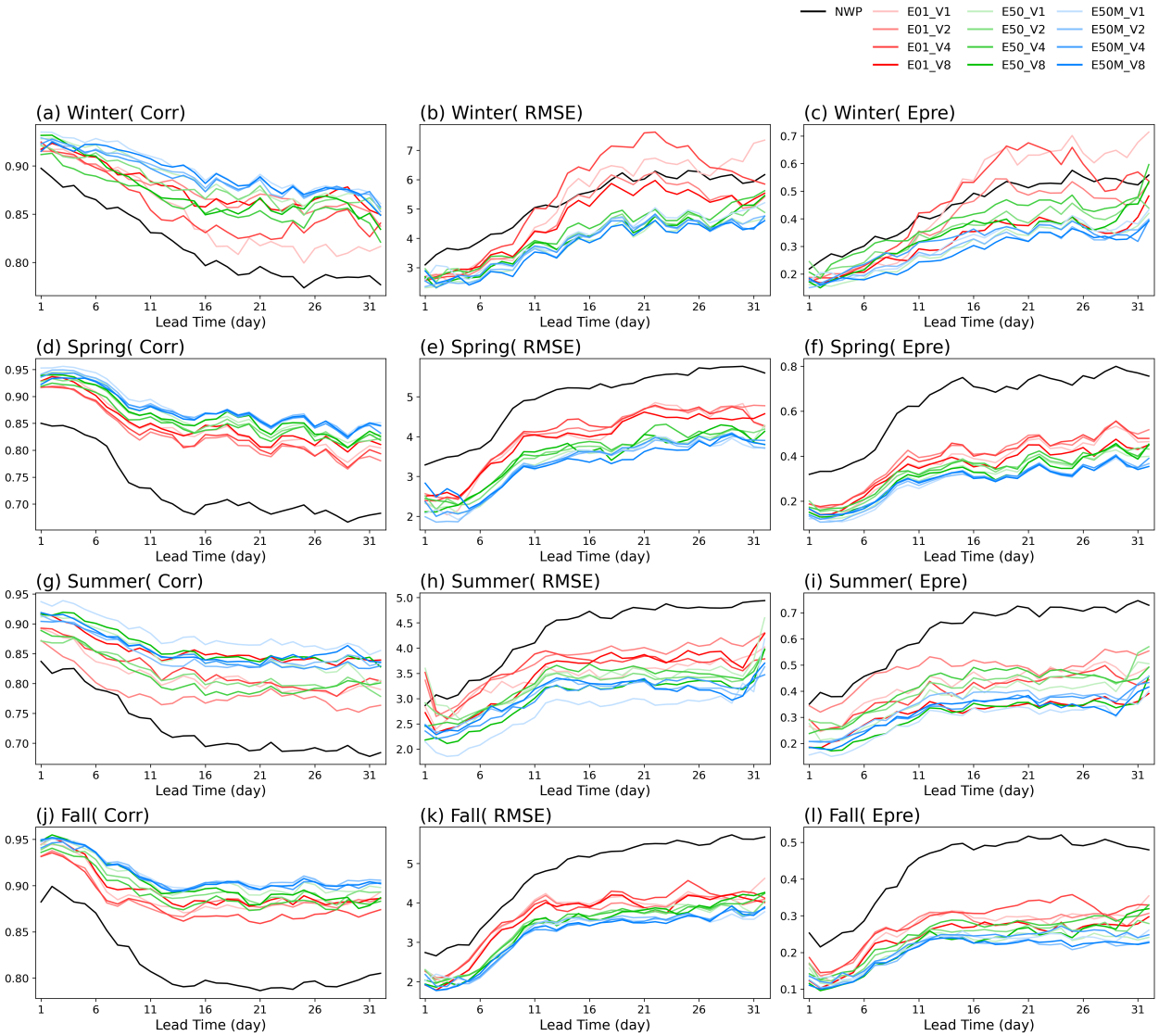

Figure R4: Figure S5) Same as Fig. R3, but for temperature forecasts.

9. I don't see the train/test procedure for E50 anywhere. There are several ways the training and testing could work, so it should be specified. This detail is essential given that this seems to be one of the main focuses of the paper.

   Thank you for your feedback. We have clarified the training and testing procedure in the revised manuscript. Specifically, ensemble members and variables are combined along the channel dimension. For example, in the experiment using all 50 ensemble members and 4 variables (E50 V4), the model receives 400 input channels. We have added this explanation and an illustrative example to the manuscript for clarity. All other training settings, including data splitting, model architecture, and optimization procedures, are kept consistent across E01, E50, and E50M configurations to ensure a fair comparison. Line 105: For example, the E50 V8 configuration has 400 input channels, while the E50M V2 configuration has 2 input channels. In the model, the input dimensions are referred to as height, width, and depth, corresponding to lead time, latitude, and longitude, with sizes of 32, 72, and 72, respectively.

10. I don't think the predictor variable sets are ever actually stated. The top 8 predictors for each target variable are given in the appendix, but this is scientifically important information that should be in the main text. It is also not clear from the text or SI which of the 8 are included in each subset (I see from the code that Vk is the k highest correlation with the target, which makes sense but is not stated).

    Thank you for constructive comment. We have moved the Table S1 from the appendix to the main text. Additionally, we have added a sentence clarifying that the Vk subset represents the top k predictors selected based on their spatial correlation with the target variable.
    Line 133: The spatial pattern correlation coefficient between the mean state of each additional variable and that of the target variable is then computed. The absolute values of these correlations are averaged across the two timescales, and the variables are ranked accordingly. Rankings are shown above each bar in Fig.S2. The top eight variables for each target are selected for use in the 3D U-Net model, as summarized in Table 2.

11. I think it is important to acknowledge somewhere in the paper that PRISM data are the output of statistical interpolation and so have uncertainties and systematic errors, which will impact the output of the NN. The paper sometimes refers to PRISM data as "observations," which is misleading.

    We agree with your point and have revised the manuscript accordingly. We have replaced references to PRISM data as "observations" with more accurate terminology and now refer to them as a "reanalysis dataset".

Minor Comments

Line 16 (and elsewhere): when you refer to a paper in running text, you should still provide the year in parentheses

Thank you for the suggestion. We have revised the in-text citations throughout the manuscript to include the publication year in parentheses when referring to papers in running text.

Line 34: Re "subset of variables," a subset of what? I think you're referring to "additional" or maybe "auxiliary" variables, but I don't think subset is the right term. If anything it's a (super)set that includes the target variables as a subset.

We agree with the reviewer that the term "subset" was imprecise. We have revised the text to clarify that we are referring to "set of additional variables". // Line 37: a broader set of additional variables

Line 36: Again I do not think "sub-variable" is the right term here. See above for suggested alternatives.

Thank you for pointing this out. We have replaced "sub-variable" with "additional variable" to maintain consistency and improve clarity.

Lines 49-51: On line 49 is says "we select forecasts from CY40R1" and on line 51 it says "We utilize forecasts from CY40R1 to CY48R1." This might be clearer to someone more familiar with ECMWF forecast naming scheme, but I find this confusing. It would be helpful to give some additional explanation what this means and maybe provide a link to the relevant data product(s).

We appreciate this comment and have revised the text to clarify that we ultimately used forecasts dataset for our experiments. We have also added a link to the ECMWF forecast model documentation to help readers unfamiliar with these cycle names. Line 57: We select the $1.5° \times 1.5°$ resolution, 50 ensemble perturbation forecasts (approximately 120 km $\times$ 120 km over the study region), twice-weekly forecast cycles, and 32-day lead times to match the earliest version of the ECMWF model (Roberts et al., 2018).
Line 61: For detailed information on each version of the model, please refer to the ECMWF model archive: https://confluence.ecmwf.int/display/S2S/ECMWF+Model.

Line 70: Remove "properly." Also, I believe the preferred GMD style is "Fig." rather than "Figure" in running text.

Thank you for feedback. "Properly" has been removed for clarity. We have also changed "Figure" to "Fig." to align with GMD style guidelines.

Line 73: I generally find it clearer to talk in terms of fine versus coarse spatial resolution.

Thank you for your helpful suggestion. In response, we have revised the description of the 3D U-Net structure to use the terms "fine" and "coarse" spatial resolution, which we agree provide greater clarity.
Line 89: The contracting path progressively reduces spatial dimensions (moving from fine to coarse) while increasing feature channels, allowing the model to capture broader contextual information. Conversely, the expanding path restores spatial resolution (from coarse to fine), enabling precise localization of weather patterns.

Line 96: Regarding "conservative interpolation," can you be more specific about the method?

Thank you for your question. We have clarified the conservative interpolation in text. This method preserves the integrated quantity over the grid area, which is particularly important for variables like precipitation.
Line 128: We then apply conservative interpolation, a method that preserves physical quantities like mass or energy during spatial grid adjustments, to ensure the accurate preservation of values during spatial adjustments.

Lines 97-98: Regarding "given the established relationship . . . ", what is the relationship?

Thank you for constructive comment. Here, "established relationship" refers to the finding that predictability can be evaluated based on the mean state, as described by Ryu et al. (2024). This sentence has been partially revised for clarity.

Line 131: Based on the fact that predictability can be evaluated using the mean state (Ryu et al., 2024), we calculate the mean state of each additional variable across both weather and sub-seasonal timescales.

Line 110: semicolon should be colon

Thank you for feedback. Corrected the semicolon to a colon as suggested.

First paragraph of 3.1: This description of the scope of analysis should come much earlier in the paper, not in the results section.

We appreciate the suggestion. However, this paragraph is intended as a brief summary of the experimental settings used for analysis, not as a statement of the overall study scope. To clarify, we have revised the paragraph accordingly.
Line 161: The performance of the 3D U-Net model, compared to traditional NWP forecasts, was evaluated across twelve cases combining three ensemble configurations and four input variable sets (Fig.S3). The 3D U-Net consistently outperformed the raw NWP forecasts across three evaluation metrics, except for E_{pre} in precipitation. Statistical tests comparing each model's evaluation metrics with those of the NWP baseline showed that, apart from the E_{pre} metric for precipitation, the improvements were significant. For precipitation E_{pre}, the results were mixed: five models (E01 V4, E50 V2, E50 V8, E50M V1, and E50M V2) showed no significant improvement, while seven models exhibited significant degradation.

Line 131: I think you mean "metrics"

Thank you for feedback. Corrected "matrices" to "metrics."

Line 141: "suggest elevation can enhance temperature post-processing accuracy" this is just lapse rate, right? I don't think this should be framed as a finding of NN methods

Thank you for pointing this out. We agree with the reviewer that the positive impact of elevation on temperature post-processing accuracy primarily reflects the lapse rate, and thus should not be considered a novel finding of neural network methods. However, incorporating elevation enables the model to better learn and correct for this effect, thereby improving performance. This result aligns with previous studies, such as Rasp and Lerch (2018), which identified elevation as the most important predictor for temperature post-processing. This sentence has been partially revised for clarity.
Line 200: This may be attributed to the inclusion of altitude, which has been shown to be one of the most important variables in temperature post-processing (Rasp and Lerch, 2018).

References

Höhlein, K., Schulz, B., Westermann, R., & Lerch, S. (2024). Postprocessing of ensemble weather forecasts using permutation-invariant neural networks. *Artificial Intelligence for the Earth Systems*, *3*(1), e230070.

Horat, N., & Lerch, S. (2024). Deep learning for postprocessing global probabilistic forecasts on subseasonal time scales. *Monthly Weather Review*, *152*(3), 667–687.

Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885–3900.

Schulz, B., & Lerch, S. (2022). Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Monthly Weather Review*, *150*(1), 235–257.