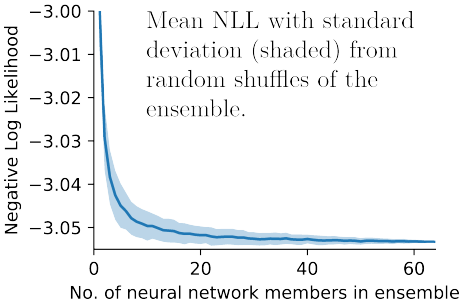
We thank the reviewers for their insightful and positive comments and the area chair for their consideration. We will address common concerns first before addressing individual reviewer comments.

**1. Methodological details.** The comments on insufficient details in the methods section are well-taken. We will add pseudocode that describes both the prior design and the training processes. **2. Novelty/relevance.** While we concede that this paper does not contain a major methodological leap, we believe it combines different elements (some borrowed from prior works) in a novel manner and lays out a roadmap for handling an important class of problems– noisy geospatial time-series data with competing physical/ empirical models of the underlying process where the task is to make forecasts or fill in large gaps. Each one of these elements is crucial to the success of the method: the pointwise-linear model combination offers immediate interpretability and makes it more palatable to domain specialists; the spatio-temporally dynamic model-weighting allows accurate predictions; the enforced quasi-periodicity in time and periodicity in space makes it extrapolate more successfully; the Bayesian framing lets us compute appropriate confidence intervals; the heteroscedastic treatment of aleatoric uncertainty accounts for the highly variable quality of available data; and the ensembling approach to Bayesian inference allows scalability. Further, given the significance of this class of problems, which includes not just climate modelling but also predicting crop yields, pollutants and disease spread, we believe it deserves the attention of the ML community broadly and not just that of domain specialists. We hope that this work can be built upon by others in the ML community and encourage cross-pollination between fields. **3. Related work.** We will be including a paragraph in the introduction which highlights recent uses of ML with climate ensembles (e.g. E. Barnes, P. Nowack, K.L. Chang). However, we believe that our work is distinctly positioned compared to these, which either lack certain considerations (no spatiotemporal weighting or uncertainty), aren't scalable, or are asking different research questions entirely. This is a young subfield, so there are few highly relevant papers. **4. Baselines.** The baseline methods are simple because these are the ones that are widely used by the community. Pure RMSE performance is not the goal here and any method that hopes to replace current standards must balance interpretability, accuracy and a good treatment of uncertainty. **5. Uncertainty evaluation.** Concerns were expressed that the uncertainty of the model has not been quantitatively evaluated. However, we clarify that in the paper we reported the fraction of data points in different subsets of the validation dataset within 1x/2x/3x of predicted standard deviations, allowing a quantitative comparison with the fractions (68.2%/95.4%/99.6%) expected within those ranges for the assumed Gaussian distribution. This point will be made more clearly and mean negative log likelihoods will also be reported.

**Reviewer 1.** Eq. 3 has been derived in the appendix since the heteroscedastic case was not considered in Pearce et al. The derivation of Eq. 2, however, is relatively lengthy, and not a contribution from us. To avoid duplication of their work, we believe it best to omit it from our paper, although we will add details on the assumptions this inference technique makes about parameters of the network. We also agree that the term "model skill" is ambiguous and unnecessary. We will remove it.

**Reviewer 2.** Responding to their numbered comments. **2)** The 'z coordinate' refers to the spatial z coordinate, which is the cosine of latitude. We agree that $x$ and $y$ are overloaded and will clean up our notation. In line 70, Euclidean spatial coordinates $(x, y, z)$ will be changed to $(u, v, w)$. **3,5)** We will add pseudocode to the methods section to further detail the prior design process and what exactly these checks are. **4)** We did not test MC (Monte Carlo) dropout since the Foong et al. paper cited makes a convincing empirical and theoretical case against it. MC dropout also performs worse compared to RMS in benchmark tests by Pearce et al. **6)** Hyperparameter optimization was mostly avoided (please see prior design section in the paper). For choice of ensemble size, please see the above plot which shows how the negative log likelihood of the test data converges as we use a larger neural network ensemble. Any ensemble greater than 30 in size would have been adequate but we ran more to ensure convergence nevertheless. **7)** The best overall model and best model at every spatial location will be added as baselines (they perform predictably worse). **Minor comments:** DU (Dobson Unit) definition will be added.




**Reviewer 3.** The space-time representation, though commonsensical, is not standard, to the best of our knowledge. This treatment of coordinates is crucial in ensuring good accuracy.

**Reviewer 4.** We believe that a stochastic variational Gaussian process is a fair comparison as the more standard GPR scales as $O(N^3)$ in the number of data points and our dataset has 1.5 million datapoints. The SVGP paper (Titsias 2009) has accrued 815 citations on Google Scholar and is certainly one of the most common ways of scaling Gaussian Processes to big data. We will cite Breimann (1996), but we observe that that regression involves a statically weighted combination of regressions. If the models are not weighted differently in different parts of the input space, the performance will be similar to the weighted mean baseline.