nature portfolio

Peer Review File

Scalable spatiotemporal prediction with Bayesian neural fields

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Authors):

Report for "Scalable Spatiotemporal prediction with Bayesian Neural Fields" (Manuscript number NCOMMS-24-09397)

Summary of the manuscript

The authors consider the problem of spatiotemporal prediction with large datasets, which is ubiquitous in many modern domains of Science and Society.
Saad et al. propose to tackle this challenging problem using Bayesian Deep Learning, as an alternative to Gaussian Processes, which provide the dominant statistical approach employed in spatiotemporal data modelling.
In particular, the authors introduce "BayesNF", a novel deep learning routine framed in a Bayesian inference setting. After describing the architecture in great detail, Saad and colleagues test its effectiveness on six publicly available, large-scale spatiotemporal datasets, and they compare its prediction accuracy against four state-of-the-art (SOTA) baselines with open-source implementation.
Finally, the statistical model employed in this work is made available via an open-source release of a software package.

Overall assessment

The manuscript deals with the important and timely problem of spatiotemporal prediction with large-scale datasets. BayesNF often convincingly outperforms SOTA baselines in many of the evaluation metrics employed (as shown extensively in Table 3 of Section 2.3), and the runtime is competitive wrt that of previously-established methods.
The work is well written, easy to follow and technically well executed. The methodology proposed by Saad et al. could find applicability in diverse scientific fields, and the inclusion of an open-source software will facilitate usability.

For the reasons mentioned above, I am positively inclined towards acceptance of the manuscript, provided that the authors can address the issues I point out in the "Major Comments" section.

Preliminary assessment of the software package "BayesNF"

I feel that the open-source repository accompanying the manuscript is one of the crucial ingredient of the submission. I did not have time to look at it in person, but I asked to one of my collaborators to provide a feedback on this specific part of the work. The following paragraph represents a summary of her first (positive) interaction with the software.

The installation process on a common MacBook was smooth for the most part. However, I encountered minor compatibility issues concerning for instance the version of numpy required (that was older than the one I have installed). To mitigate such minor problems for future users, I would suggest the inclusion of a virtual environment (e.g. conda, venv) with compatible versions of libraries utilized in the project.

The tutorials proved to be a good resource, allowing to implement the model using the free version of Google Colab. Given the computational limitations, I had to use a small ensemble size (3), and a thousand training epochs. Despite these adjustments, the predictions obtained were satisfactory, testifying the robustness of the proposed model. Finally, I would like to point out a typo in the figures produced by the London Air Quality tutorial. On the "y" axis in the plot the label is "flu cases".

Major Comments

1. The proposed statistical model for spatiotemporal prediction incorporates a set of hyperparameters that the authors define in Section 2.1 and Listing 1. These include: (i) the number of hidden layers H (which in BayesNF is fixed to H=2, if I understand correctly); (ii) the size of the hidden layers N_d (which is never specified throughout the manuscript) (iii) the variance of the priors over the parameters at each layer (are these prior Gaussian? This is never clearly stated in Section 2.1 or in the Methods, but I could have missed it).

   I would expect that the prediction error could be significantly affected by the value of these hyperparameters, but the manuscript does not address this important point. (i) Did the authors try to employ architectures with H larger than 2? (ii) How does the prediction error change increasing the hidden layers size N_d? (iii) Does the magnitude of the variance of the prior over the parameters at the each layer influence the prediction error?

   For instance, theoretical and empirical work on the infinite-width limit of Bayesian deep neural networks [1-3] suggests that wide networks usually outperform finite-width networks in the case of fully-connected architectures (at least in standard computer vision tasks), while a small variance for the prior at the last hidden layer should be beneficial for generalization.

   While I understand that a systematic study of the optimal hyperparameters choice is computationally very demanding, I would ask the authors to perform a preliminary analysis in this direction, and to include that in a Supplemental Material.

2. BayesNF is essentially a Bayesian fully-connected architecture with three layers, which takes as input the spatiotemporal covariates defined in Eqs. (5-8). Two additional extra-steps enter the definition of the model: (i) the covariate scaling layer; (ii) the convex combination layer, which learns an effective activation function.

   The authors only briefly explain in Section 2.1 why these choices are made. I would find useful and instructive to add a section in the supplemental material that describes the practical benefits of including (i) and (ii) in BayesNF, possibly comparing the generalization performance (prediction error) of BayesNF with that of a vanilla Bayesian network that does not include extra-steps (i) and (ii) (on at least one of the spatiotemporal datasets they already employ in the main text).

Minor Comments

1. I personally appreciate the visual impact of Fig. 5 of the manuscript, but I would add numerical values and units of measurement on the axes.
2. I found Section 2.2 somewhat difficult to follow and too technical (especially the subsection “variography” and Fig.3). I would move that after the current section 2.3 or in the Supplemental material.
3. The authors employ the same letter “d” to indicate the dimension of the spatial coordinates (see first paragraph of Section 2.1) and the labelling of the hidden layers in Listing 1 (and in the “Hidden layers” paragraph of Section 2.1). I find this potentially confusing and I would suggest the authors to use a different letter for the hidden layers labelling.

References

[1] J. Lee et al.; *Deep Neural networks as Gaussian Processes*; ICLR (2018).

[2] R. Novak et al.; *Bayesian deep convolutional networks with many channels are Gaussian Processes*; ICLR (2018).

[3] J. Lee et al.; *Finite versus Infinite neural networks: an empirical study*; Neurips (2020).

Reviewer #2 (Remarks to the Authors):

This paper introduces an approach to learning predictive models for spatiotemporal data based on a Bayesian neural network (BNN) framework. The approach, named Bayesian Neural Fields (BayesNF), is tested on a variety of problems in climate, weather, and epidemiology prediction, with performance improvements shown over sparse variational GPs, gradient boosting trees, generalized linear models, and trend surface regression.

Overall my feelings about the paper are somewhat mixed. On the one hand, the high-level suggested approach seems sensible, the method appears to perform well, good quality code is provided, and I like that it is attacking an important problem. On the other hand, I do not feel that the paper makes a very strong case for the precise model being used and its novelty, I do not feel the paper sufficiently discusses related work, there are elements of the presentation that could be improved, and I have some concerns with the experimental comparisons and the competitiveness of the chosen baselines. A quick google scholar search shows that there are various other BNN or Bayesian graph NN methods for spatiotemporal data (e.g. [1-5]), including some that are not cited, and at present, I do not think the paper makes a strong enough case for its novelty and significance relative to these. My expertise is very much more on the machine learning / BNN side of the work than spatiotemporal applications so my concerns may be due to my lack of familiarity, but I find it concerning that there are no other deep learning-based baselines considered in the experiments or reasonably discussed in related work. At the very least, I think the paper needs significantly more discussion of related probabilistic deep learning approaches to the problem and possible alternatives (e.g. neural processes), rather than just having a predominant focus on Gaussian process approaches as is the case at present.

In light of the concerns above, I do not feel able to back acceptance of the paper in its current form. However, there are aspects of the work I like and so I would potential be supportive with appropriate updates if the authors can better convince me of the novelty and utility of the specific approach relative to other deep learning approaches.

Specific comments (here I see 1-4 as major concerns that would be potential blockers to acceptance, while 5-10 are either more minor or more question based)

1. While most of the low-level writing is good and clear, there are some high-level aspects of the paper that I think significantly detract from its clarity.

First, I the method is immediately presented in a much more technical and low-level algorithm manner than is needed. Very little intuition or description of the key ideas / features of the BayesNF are provided when describing the method, with the exposition instead immediately launching into the exact architecture will little context. Improving this would not only assist with the clarity, but also help make the novelty of the approach clearer.

Second, I think some of the mathematical descriptions are also much heavier than they need to be, and I think the paper will not be very accessible to those outside the machine learning field (with the paper presumably intended to appeal to such an audience, given it was not submitted to a ML venue).

Third, I believe the structuring of the experiments makes the utility of the approach more difficult to establish. Namely, it is very difficult for the reader to garner much intuition from the qualitative assessments of the method as there is no reference point or even much in the way of expected behavior. I found myself wanting to just skip ahead to the quantitative evaluations.

2. I found that some of the claims about BayesNF were a little strong, such as the fact that it provides "robust uncertainty quantification" and is "well-calibrated". To be clear, I do not think this is a paper that is egregiously overclaiming, but I do think it currently make clear some of the limitations with the framework being used and some of the claims might mislead non-expert readers on this. In particular, I think t is important for the paper to make clear the well-documented shortcomings of BNNs and very approximate nature of the inference schemes being used. For example, readers who are not familiar with BNNs are currently liable to miss the fact

that approximate "posteriors" being derived are actually likely to be very far from the true posterior, or that sequential updating of these BNNs schemes tends to be very far from satisfying Bayesian consistency. An explicit discussion on the limitations of the approach and BNNs more generally would be very helpful here. The paper currently also has very little discussion of any kind of calibration that is being applied, but I would assume that at least some is happening (e.g. through tuning the KL scaling factor when doing VI)?

3. As discussed earlier, I believe that it is necessary to have some more heavy duty baselines compared to for the quantitative experiments. Quite a few of the current comparisons are either very old or very simple methods, and no other deep learning approaches are considered at all. I’m not familiar enough with the application area to say exactly which baselines should be considered, but at a very least if there are no suitable specific methods (which I think it rather unlikely) I would expect to see some kind of naive BNN baseline or some other more modern method like a neural process.

4. I don’t think that it is acceptable to just report the “better of MAP or variational inference” when reporting the BayesNF results in section 2.3. Using different inference schemes here equates to different methods (noting that they are both far from the true posterior) and this is pretty blatantly biasing the results in the BayesNF’s favour. The results in Figure 6 also later show that VI, while usually preferable, appears to occasionally have catastrophic failures. I think that the presentation in section 2.3 is somewhat hiding these at the moment and this needs more careful highlighting and discussion. I would suggest that both inference methods are included in Table 3.

5. I felt the discussion around focusing on low frequencies to be a bit over simplified and lacking in sufficient nuance. Low frequency signal elements are often highly desirable for extrapolation and capturing long term trend, while you don't want overly high frequency aspects either as these will often correspond to noise fluctuations. I think the underlying arguments here are fine, but I think it needs explaining more carefully.

6. In figure 2b it looks to me like there are some axis-aligned artefacts that are unlikely to reflect true underlying behaviour, e.g. there are often long thin strips where the prediction is very consistent. I would be interested to see ho the prediction changes if the input coordinates were instead rotated. This shouldn't substantially change the predictions but I suspect it will and this could indicate some shortfalls in the approach.

7. I did not find the results in the variography sufficient to fully evidence the quite strong conclusions reached. Comparing the plots quantitively is difficult and there is not even a naive baseline for what can of fit might be expected. From my understanding this is also not a infallible comparison either: it was not clear if this is based on held out test data or not, how much tuning there had been to get a good fit, etc. The discrepancy at the 0 and 1 day lags also seems to be quite large but was never explained. It is quite possible that I have misunderstood the results here, but at the very least this needs explaining more carefully to an audience not familiar with the particular problem, and it may be that the claims also need to be tuned down.

8. It was unclear if there is fully held out test data in the quantitative experiments or whether this had been used during the model development (and or hyperparameter tuning).

9. It wasn’t clear to me why the approach described itself as doing “hierarchical inference”

10. I do not think that Table 1 adds much and would suggest removing it from the main paper.

References

[1] Gao, Fang, Zidong Xu, and Linfei Yin. "Bayesian deep neural networks for spatio-temporal probabilistic optimal power flow with multi-source renewable energy." Applied Energy 353 (2024): 122106.

[2] Wang, Jian, et al. "Predicting wind-caused floater intrusion risk for overhead contact lines based on Bayesian neural network with spatiotemporal correlation analysis." Reliability Engineering

& System Safety 225 (2022): 108603.

[3] Liu, Yongqi, et al. "Probabilistic spatiotemporal wind speed forecasting based on a variational Bayesian deep learning model." Applied Energy 260 (2020): 114259.

[4] Xia, Jiangnan, et al. "Multi-view Bayesian spatio-temporal graph neural networks for reliable traffic flow prediction." International Journal of Machine Learning and Cybernetics 15.1 (2024): 65-78.

[5] McDermott, Patrick L., and Christopher K. Wikle. "Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data." Entropy 21.2 (2019): 184.

Reviewer #2 (Remarks on code availability):

The code all appears to be clean, clear, and well documented. I have not run it directly myself.

Author Response

We thank Reviewers 1 and 2 for their thoughtful comments about our work and are very grateful for their feedback and suggestions to improve the manuscript. After carefully reading the reviews, we have written several clarifications and added new experimental evaluations to the manuscript. We believe these changes greatly improve the paper and address all the main points in the two reviews.

A table summarizing the main changes in the manuscript is provided at the end of this document. We next provide a point-by-point response to the reviewer's comments, providing clarifications and references to locations in the revised manuscript where changes have been made. New or revised text in the manuscript is shown in red font.

Reviewer 1

The proposed statistical model for spatiotemporal prediction incorporates a set of hyperparameters that the authors define in Section 2.1 and Listing 1. These include: (i) the number of hidden layers H (which in BayesNF is fixed to H=2, if I understand correctly); (ii) the size of the hidden layers Nd (which is never specified throughout the manuscript) … two additional extra-steps enter the definition of the model: (i) the covariate scaling layer; (ii) the convex combination layer, which learns an effective activation function. The authors only briefly explain in Section 2.1 why these choices are made. I would find useful and instructive to add a section in the supplemental material that describes the practical benefits of including (i) and (ii) in BayesNF, possibly comparing the generalization performance (prediction error) of BayesNF with that of a vanilla Bayesian network that does not include extra-steps (i) and (ii)

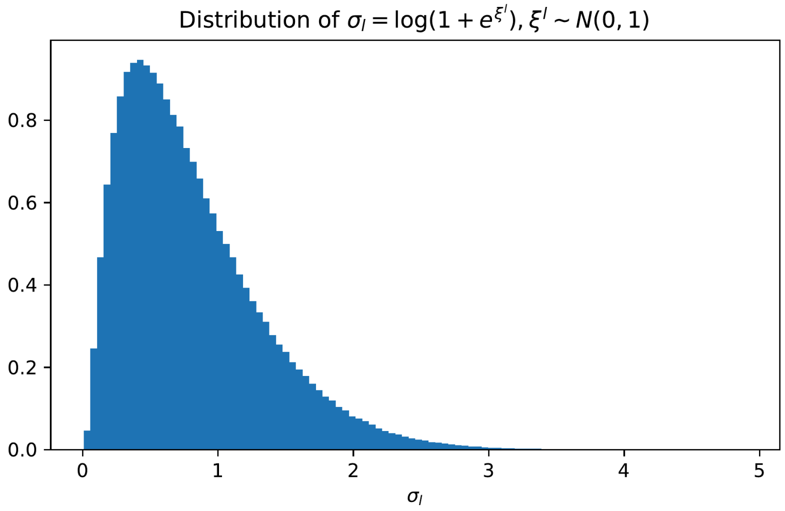
To address these suggestions, we have added a new "Section 4.5.2: Model Architectures" and Figure 7 (panels (a)–(h)) on Pages 20–24 of the revised manuscript. We reran all six benchmark experiments and computed the new prediction errors (RMSE, MAE, MIS) and wall-clock runtime using the original version of the model (MAP inference; 64 particles; fixed number of training epochs) while applying a single change to the network. The considered changes in the ablation study are:

- Halving the width of the hidden layers	(Fig 7a)
- Doubling the width of the hidden layers	(Fig 7b)
- Increasing network depth by 1	(Fig 7c)
- Decreasing network depth by 1	(Fig 7d)
- No convex combination layer (only tanh activation)	(Fig 7e)
- No convex combination layer (only elu activation)	(Fig 7f)
- No covariate scaling layer	(Fig 7g)
- No spatial Fourier features	(Fig 7h)

The new Section 4.5.2 discusses in detail how each of these modeling ablations influences the generalization performance and runtime. We have also updated the online tarballs to include predictions produced by all of the above ablations.

(iii) the variance of the priors over the parameters at each layer (are these prior Gaussian? This is never clearly stated in Section 2.1 or in the Methods, but I could have missed it). (iii) Does the magnitude of the variance of the prior over the parameters at the each layer influence the prediction error? For instance, theoretical and empirical work on the infinite-width limit of Bayesian deep neural networks [1-3] suggests that wide networks usually outperform finite-width networks in the case of fully-connected architectures (at least in standard computer vision tasks), while a small variance for the prior at the last hidden layer should be beneficial for generalization.

The variance of the normal prior over parameters at each layer is not a fixed hyperparameter but rather a *learnable parameter*. Each layer *l* = 1, …, *L* has its own variance. On Page 7, the red text in Listing 1 expresses the prior as σ□ = ln(1 + exp(ξ^l)) for ξ^l ~ N(0,1). We have further clarified this point in the "Hidden Layers" section on Page 7. The implied prior over σ□ is shown below.



As the reviewer notes, the variance can play an important role in the generalization error. However, as it is typically not obvious what the correct variance to use is, BayesNF specifies the variance as a learnable parameter rather than a fixed hyperparameter.

I personally appreciate the visual impact of Fig. 5 of the manuscript, but I would add numerical values and units of measurement on the axes.

This figure (now Fig. 3 on Page 13 of the revised manuscript) has been updated to include numerical values and units of measurements. The units have also been added in the *Datasets* subsection on Pages 8–9.

I found Section 2.2 somewhat difficult to follow and too technical (especially the subsection “variography” and Fig.3). I would move that after the current section 2.3 or in the Supplemental material.

The order of Section 2.2 and Section 2.3 in the original has been switched, as suggested.

The authors employ the same letter "d" to indicate the dimension of the spatial coordinates (see first paragraph of Section 2.1) and the labelling of the hidden layers in Listing 1 (and in the "Hidden layers" paragraph of Section 2.1). I find this potentially confusing and I would suggest the authors to use a different letter for the hidden layers labelling.

We have changed the labelling of the hidden layers from *d* to *l* (and for further clarity, the number of hidden layers from *H* to *L*). These changes appear on Page 7 (Listing 1 and main text) of the revised manuscript.

The installation process on a common MacBook was smooth for the most part. However, I encountered minor compatibility issues concerning for instance the version of numpy required (that was older than the one I have installed). To mitigate such minor problems for future users, I would suggest the inclusion of a virtual environment (e.g. conda, venv) with compatible versions of libraries utilized in the project.

We have added the compatible versions of libraries using Python 3.10, which is used by the Github Actions integration test, to the repository. We have also added instructions for how to create a fresh Python virtual environment to install the software.

https://github.com/google/bayesnf/commit/d56cb32613fe7395e04826e9050d1cd73b464236

Finally, I would like to point out a typo in the figures produced by the London Air Quality tutorial. On the "y" axis in the plot the label is "flu cases".

The online tutorial has been fixed and republished.
- https://github.com/google/bayesnf/pull/44
- https://google.github.io/bayesnf/tutorials/BayesNF_Tutorial_on_London_Air_Quality/

Reviewer 2

1. While most of the low-level writing is good and clear, there are some high-level aspects of the paper that I think significantly detract from its clarity.

First, I the method is immediately presented in a much more technical and low-level algorithm manner than is needed. Very little intuition or description of the key ideas / features of the BayesNF are provided when describing the method, with the exposition instead immediately launching into the exact architecture will little context. Improving this would not only assist with the clarity, but also help make the novelty of the approach clearer.

To improve the clarity we have added a high-level introduction to the key stages of BayesNF when describing Figure 1, shown in Lines 171–186 of Page 6 in the revised manuscript. This description explains the key architectural choices and the modeling capabilities that they aim to achieve before diving into the technical details.

Second, I think some of the mathematical descriptions are also much heavier than they need to be, and I think the paper will not be very accessible to those outside the machine learning field (with the paper presumably intended to appeal to such an audience, given it was not submitted to a ML venue).

We hope the above changes on Page 6 now strike a balance between high-level descriptions and technically precise descriptions needed for practitioners to understand, reproduce, and build on our method. We are happy to work with the Editorial team to determine whether more of the technical material should be moved to Section 4: Methods and/or the Supplementary.

Third, I believe the structuring of the experiments makes the utility of the approach more difficult to establish. Namely, it is very difficult for the reader to garner much intuition from the qualitative assessments of the method as there is no reference point or even much in the way of expected behavior. I found myself wanting to just skip ahead to the quantitative evaluations.

Reviewer 1 also noted it would be helpful to switch the order of the quantitative benchmark (Section 2.2; Page 8) and qualitative data analysis (Section 2.3; Page 14) . We have implemented this change and hope it helps give the readers a better reference point.

3. As discussed earlier, I believe that it is necessary to have some more heavy duty baselines compared to for the quantitative experiments. Quite a few of the current comparisons are either very old or very simple methods, and no other deep learning approaches are considered at all.

We appreciate the reviewer's request for implementing more "heavy duty" baselines. To this end, we have added the "Neural Basis Expansion Analysis" (NBEATS; Oreshkin et al. 2019) baseline to the quantitative evaluation. Page 9 of the revised manuscript describe NBEATS. Pages 11–12 discuss the results using this method. We have also updated the library with all the scripts to run the benchmarks using the NBEATS method (and more generally, any method from the NeuralForecast package).

Regarding the claim "Quite a few of the current comparisons are either very old or very simple methods", we wish to clarify that the ST-SVGP and GLMM baselines are both new and highly competitive, in particular:

> (a) The Sparse SpatioTemporal Variational Gaussian Process (ST-SVGP) is a new method from Hamelijnck et al. (NeurIPS 2021). It integrates many sophisticated techniques that are needed for GPs to scale to large spatiotemporal datasets, such as the stochastic PDE formalism; natural gradient variational inference; and parallel Bayesian filtering and smoothing on GPUs using JAX. Hamelijnck et al. 2021 show that ST-SVGP is superior to previous sparse GPs such as KISS-GP (Wilson and Nickish, ICML 2015) and the vanilla SVGP (Titsias, AISTATS 2009).
>
> (b) The Spatiotemporal Generalized Mixed Effect Model (ST-GLMM) we used also incorporates very new methods from Anderson et al. bioRxiv 2022 that enable accuracy and scalability of these methods, namely Gaussian-Markov Random Fields and stochastic PDEs. The library has been updated as recently as 2024-04-03 and enjoys a broad user base with many compelling real-world spatiotemporal case studies and numerical evaluations.

To the best of our knowledge, GP/GLMM-based methods remain state-of-the-art for probabilistic prediction in sparse, large-scale spatiotemporal data. They are workhorse models used by practicing statisticians and continue to be the focus of very new textbooks, such as Paula Moraga's 2023 book *Spatial Statistics for Data Science with: Theory and Practice with R* and Wikle et al. 2019 book *Spatio-Temporal Statistics with R*. The evaluation results using the NBEATS baseline (via the NeuralForecast package) sheds further light on this claim: it required an expensive hyperparameter search and still did not generally deliver competitive results. We also tried other NeuralForecast models, including *AutoFormer*, *Informer*, *DeepAR*, *TemporalFusionTransformer*, and *TiDE*, but these methods did not outperform NBEATS and so we do not report their results.

Ultimately, we hope our new benchmark datasets make it easier for researchers to develop and evaluate new deep-learning based methods for the challenging class of spatiotemporal probabilistic prediction problems considered in this work.

I'm not familiar enough with the application area to say exactly which baselines should be considered, but at a very least if there are no suitable specific methods (which I think it rather unlikely) I would expect to see some kind of naive BNN baseline or some other more modern method like a neural process.

In addition to the NBEATS baseline, we have added a new "Section 4.5.2: Model Architectures" and Figure 7 (panels (a)–(h)) on Pages 20–24. The last four of these ablation studies can be understood as various naive BNN baselines which remove key modeling choices from BayesNF. A summary of the ablation results and discussion of the benefits of the BayesNF architecture are given on Page 21–22.

A quick google scholar search shows that there are various other BNN or Bayesian graph NN methods for spatiotemporal data (e.g. [1-5]), including some that are not cited, and at present, I do not think the paper makes a strong enough case for its novelty and significance relative to these. My expertise is very much more on the machine learning / BNN side of the work than spatiotemporal applications so my concerns may be due to my lack of familiarity, but I find it concerning that there are no other deep learning-based baselines considered in the experiments or reasonably discussed in related work. At the very least, I think the paper needs significantly more discussion of related probabilistic deep learning approaches to the problem and possible alternatives (e.g. neural processes)

We thank the reviewer for emphasizing the need to better discuss Related Work. We have added "Related Work" on Pages 3–4 of the revised manuscript and citation and discussion of all these methods. To recap the main discussion:

Neural processes (Garnelo et al. 2019) integrate deep neural networks with probabilistic modeling, but are based on a graphical model structure that is fundamentally difficult to apply to spatiotemporal datasets. In particular, because neural processes aim to "meta-learn" a prior distribution over random functions, the authors note it is essential to have access to a large number of independent and identically distributed (i.i.d.) datasets during training, writing: "*To learn such a distribution over random functions, rather than a single function, it is essential to train the system using multiple datasets concurrently, with each dataset being a sequence of inputs x1:n and outputs y1:n.*" However, most spatiotemporal data analyses are based on only a single real-world dataset (e.g., the 6 benchmarks in Table 1 of Page 10 of the manuscript) where there is no notion of sharing statistical strength across multiple i.i.d. observations of the entire field. Therefore, Neural Processes are not really applicable to our evaluation.

Graph neural networks (GNNs), surveyed in Ming et al. 2023, are another popular deep-learning approach for spatiotemporal prediction which have been particularly useful in settings such as analyzing traffic or population-migration patterns. These models require as input a graph describing the connectivity structure of the spatial locations, which makes them less appropriate for spatial data that lack such discrete connectivity structure. Moreover, the requirement that the graph be fixed makes it harder for GNNs to interpolate or extrapolate to locations that are not included in the graph at training time. BayesNF, on the other hand, operates over continuous space, and is therefore more appropriate for spatial data without known discrete connectivity structure. In addition, as noted in Ming 2023, GNNs have not yet been demonstrated on probabilistic prediction tasks, and we are unaware of the existence of open-source software libraries based on GNNs that can easily handle the sparse or continuous-domain datasets in our evaluation.

Regarding the other Bayesian methods for spatiotemporal data (e.g. refs [1-5] from the review), we wish to note that these methods are highly task-oriented; e.g., [1] considers optimal power flow analysis; [2] considers wind-caused floater intrusion risk for overhead contact lines; [3] considers wind speed forecasting; and [4] considers traffic flow prediction. A closer look at these papers shows that the network architectures are carefully designed for the analysis problem at hand (e.g., power flow equations in [1]). These works also do not aim to provide software libraries that are easy for practitioners to apply in new spatiotemporal datasets outside of the application domain. In contrast, a central goal of BayesNF is to provide a domain-general modeling tool that is easily applicable to the same type of spatiotemporal datasets as Gaussian process or GLM regression models, without the need to redesign substantial parts of the probabilistic model or network architecture for each new task.

2. I found that some of the claims about BayesNF were a little strong, such as the fact that it provides "robust uncertainty quantification" and is "well-calibrated". To be clear, I do not think this is a paper that is egregiously overclaiming, but I do think it currently make clear some of the limitations with the framework being used and some of the claims might mislead non-expert readers on this. In particular, I think t is important for the paper to make clear the well-documented shortcomings of BNNs and very approximate nature of the inference schemes being used. For example, readers who are not familiar with BNNs are currently liable to miss the fact that approximate "posteriors" being derived are actually likely to be very far from the true posterior, or that sequential updating of these BNNs schemes tends to be very far from satisfying Bayesian consistency. An explicit discussion on the limitations of the approach and BNNs more generally would be very helpful here.

When we say that BayesNF provides "robust uncertainty quantification" and is "well-calibrated", we mean to say that the *predictions* produced by BayesNF have these properties. We have added the word "*predictive* uncertainty quantification" to the Abstract (Page 1).

We support these claims by evaluating predictions on benchmark data not only for point forecasts (as measured by RMSE and MAE) but also for the 95% interval forecasts (as measured by MIS). The MIS—which is the "Mean Interval Score" shown in Equation (26) on Page 19 of the revised manuscript —measures how well the 95% prediction interval (l_i, u_i) is calibrated around the true value y_i. It penalizes the width of the interval u_i - l_i and cases where the true data point y_i lies outside the prediction interval. The MIS error from BayesNF is consistently lower than the baselines, which is shown quantitatively on Page 12 and qualitatively on Page 13 of the revised manuscript.

We are in complete agreement with the reviewer that the approximate posteriors being derived may be far from the true posterior, especially over network parameters. However, we do not believe this gap to pose a big problem in practice for two reasons. First, the empirical results show that the predictive uncertainty in BayesNF is more accurate than current baselines, even with approximate Bayesian inference. Moreover, Figure 6 shows that quantifying uncertainty using MAP or VI ensembles is almost always better than maximum-likelihood estimation (MLE), which ignores the parameter priors. Second, as BayesNF is a deep neural network, the posterior distributions over parameter values such as network weights and biases may not be of inherent interest to a practitioner in a given spatiotemporal data analysis application. Rather, we expect BayesNF to be primarily used in cases where the predictive calibration is more relevant to the user.

We have added these points to Pages 17 (Discussion section) of the revised manuscript.

Regarding the model's key limitations and ideas for future improvements, these are discussed in the final paragraph of Section 3: Discussion (Pages 17–18).

4. I don’t think that it is acceptable to just report the “better of MAP or variational inference” when reporting the BayesNF results in section 2.3. Using different inference schemes here equates to different methods (noting that they are both far from the true posterior) and this is pretty blatantly biasing the results in the BayesNF’s favour. The results in Figure 6 also later show that VI, while usually preferable, appears to occasionally have catastrophic failures. I think that the presentation in section 2.3 is somewhat hiding these at the moment and this needs more careful highlighting and discussion. I would suggest that both inference methods are included in Table 3.

We have updated the Evaluation table on Page 12 of the revised manuscript to show separate entries for BayesNF VI and BayesNF MAP. The previous ablation for BayesNF without spatial features has been moved to the new Section 4.5.2. In one benchmark (Sea Surface Temperature) BayesNF MAP is the stronger method; and in one other benchmark (Precipitation) there is a tie between BayesNF MAP and BayesNF VI (in particular, the difference in RMSE, MAE, and MIS values using MAP and VI are not statistically significant; however they are both statistically significantly lower than the next-best baseline). We further discuss these points on Page 11 of the revised manuscript.

To align the results, we re-ran the evaluation from scratch using BayesNF VI and MAP with the same inference configuration (# particles and # epochs) across all benchmarks, which explains the increase in runtime and reduction in error for several BayesNF entries in Table 1. These changes do not alter the conclusions. We also wish to note the original manuscript had a transcription error for the MAE of

ST-SVGP in Air Quality 1—the original value was 2.91 but the correct value is 3.91, which has been updated in the revised manuscript.

5. I felt the discussion around focusing on low frequencies to be a bit over simplified and lacking in sufficient nuance. Low frequency signal elements are often highly desirable for extrapolation and capturing long term trend, while you don't want overly high frequency aspects either as these will often correspond to noise fluctuations. I think the underlying arguments here are fine, but I think it needs explaining more carefully.

Page 3 of the revised manuscript has been updated to clarify the role of the frequencies. We agree that using overly high frequencies can potentially make the neural networks fit noise fluctuations. However, in BayesNF many parameters at the observation layer—such as Θ_{y,1} in Eq (10) (the Gaussian variance) or Θ_{y,1,2} in Eq (11) (the StudentT scale/dof)—are not input-specific but rather shared across all inputs, which aims to mitigate the model's sensitivity to these noise fluctuations. Page 8 of the revised manuscript now includes this discussion (red text after Eqs (10)–(13)).

6. In figure 2b it looks to me like there are some axis-aligned artefacts that are unlikely to reflect true underlying behaviour, e.g. there are often long thin strips where the prediction is very consistent.

We agree that Figure 2b contains axis-aligned artifacts where predictions are consistent along certain thin regions, which are a result of the spatial Fourier features. How closely these artifacts reflect the true behavior could be further investigated by obtaining PM10 measurements at the novel locations along these regions. We have added this point on Page 16 of the revised manuscript.

7. I did not find the results in the variography sufficient to fully evidence the quite strong conclusions reached. Comparing the plots quantitively is difficult and there is not even a naive baseline for what can of fit might be expected. From my understanding this is also not a infallible comparison either: it was not clear if this is based on held out test data or not, how much tuning there had been to get a good fit, etc. The discrepancy at the 0 and 1 day lags also seems to be quite large but was never explained. It is quite possible that I have misunderstood the results here, but at the very least this needs explaining more carefully to an audience not familiar with the particular problem, and it may be that the claims also need to be tuned down.

Variography is a key tool in geostatistical data analysis. Figure 5 compares the empirical variogram computed using spatial locations at which real data is measured with the inferred variogram computed using simulations from BayesNF at novel spatial locations at which no data exists. The idea here is to see how well BayesNF extrapolates the covariance structure from the observed locations to novel locations. The ability of BayesNF to learn a predictive distribution at novel locations is a distinguishing feature of the model as compared to say Graph Neural Networks or the Bayesian recurrent neural network model of McDermot and Wikle (citation [5] in your original review), which operate over a fixed set of spatial locations. For this data analysis task there is no notion of "held-out test data". We agree that the comparison is not infallible, primarily because the true distribution at novel locations is always unknown. However, it is common practice for analysts to construct models that capture some form of regularity or continuity in space and time: in absence of any such assumption, it would be impossible to make generalizations from observed data to unobserved data.

The discrepancy at 0 and 1 day lags is discussed in the original manuscript: "*The difference between the semi-variograms is highest for tau in {0,1,2} days, suggesting that the learned model is expressing relatively smooth phenomena and assuming that the high-frequency day-to-day variance is due to unpredictable independent noise*." (Page 17 of the revised manuscript.)

In the revised manuscript, we have also tuned down the claims by applying the following changes on Pages 16–17:

- "we can **assess->gain more insight into** how well the learned spatiotemporal field matches…"
- "**confirms->suggests** that BayesNF captures these longer-term temporal dependencies."
- "**confirms->suggests** that BayesNF accurately generalizes …"

9. It wasn't clear to me why the approach described itself as doing "hierarchical inference"

The approach leverages "inference in a hierarchical Bayesian model", where the generative model is shown in Listing 1.

10. I do not think that Table 1 adds much and would suggest removing it from the main paper.

As requested, the previous Table 1 has been moved out of Section 2 to the end of the paper.

Summary of Revisions

Reviewer #1

Request	Change
Perform ablation studies showing how the prediction error on benchmark data varies with the following model changes: (i) number of hidden layers; (ii) size of the hidden layers; (iii) covariate scaling layer; (iv) convex combination layer.	All these ablation results are provided in Figure 7 panels (a)–(h) and discussed in the new Section 4.5.2 on Pages 20–24 of the revised manuscript. We have also updated the online artifact with all the ablation results.
Clarify the role/sensitivity of the variance of the prior over parameters in each layer.	The prior variance in BayesNF is a learnable parameter not a fixed hyperparameter. We have emphasized this point on Page 7 (Listing 1 and main text) of the revised manuscript.
Add numerical values and units to Figure 5	The figure has been updated accordingly (now Figure 3 on Page 13).
Switch order of Section 2.2 and Section 2.3	The sections have been switched on Pages 8 and 14.
Letter "d" is used for both dimensionality and for indexing the depth of the network layer.	The overloading has been fixed by replacing letter *d* with letter *l* for indexing the layer depth on Page 7 (Listing 1 and main text).
Include a virtual environment (e.g. conda, venv) with compatible versions of libraries utilized in the project.	The open-source library has been updated with the compatible library versions used in the integration test and instructions for installing it into a fresh venv.
Typo in the figures produced by the London Air Quality tutorial online.	The typo has been fixed and the online tutorial republished.

Reviewer #2

Request	Change
Very little intuition or description of the key ideas / features of the BayesNF are provided when describing the method, with the exposition instead immediately launching into the exact architecture will little context	On Page 6 of the revised manuscript, the key ideas and features of the BayesNF have been described in greater detail to give the reader more context before launching into the exact architecture.
I believe the structuring of the experiments makes the utility of the approach more difficult to establish… I found myself wanting to just skip ahead to the quantitative evaluations.	Following suggestions from both reviewers, the quantitative experiments now appear before the qualitative experiments by switching Sections 2.2 and 2.3 (Pages 8 and 14 in the revised manuscript).

I believe that it is necessary to have some more heavy duty baselines compared to for the quantitative experiments.	The Evaluation table on Page 12 has been updated with the Neural Basis Expansion Analysis (NBEATS) baseline from the NeuralForecast package. Page 9 introduces NBEATS and Pages 11–12 discuss the results.
I would expect to see some kind of naive BNN baseline	Ablation studies in the new Section 4.5.2 on Pages 20–24, specifically the last 4 panels of Figure 7, provide naive BNN baselines that remove architectural choices from BayesNF, showing how the errors become larger in several important cases.
I find it concerning that there are no other deep learning-based baselines … reasonably discussed in related work.	On Pages 3–4, a new Related Work section cites and discusses (1) Neural Processes; (2) Graph Neural Networks; and (3) additional domain-specific Bayesian spatiotemporal models [1-5] from R2's original review.
I found that some of the claims about BayesNF were a little strong, such as the fact that it provides "robust uncertainty quantification" and is "well-calibrated".	The Abstract (Page 1) and Discussion (Page 17) now clarify that "robust uncertainty quantification" and "well-calibrated" is over the predictions, not parameters, i.e., as measured by Mean Interval Score improvements over baselines.
Readers who are not familiar with BNNs are currently liable to miss the fact that approximate "posteriors" being derived are actually likely to be very far from the true posterior,	The Discussion on Page 17 of the revised manuscript emphasizes the posteriors are fundamentally approximate and are not guaranteed to match the true Bayesian posterior.
I don't think that it is acceptable to just report the "better of MAP or variational inference" when reporting the BayesNF results in section 2.3.	The Evaluation table Page 12 of the revised manuscript shows separate entries for BayesNF VI and BayesNF MAP. These are discussed on Page 11.
I felt the discussion around focusing on low frequencies to be a bit over simplified and lacking in sufficient nuance.	Pages 3 and 8 of the revised manuscript have been updated to better discuss low and high frequencies as well as their role in the BayesNF model.
In figure 2b it looks to me like there are some axis-aligned artefacts	Page 14 of the revised manuscript discusses these axis-aligned artifacts, which arise from the spatial Fourier features.
I did not find the results in the variography sufficient to fully evidence the quite strong conclusions reached.	On Page 14 of the revised manuscript, the wording has been updated to clarify and weaken the claims regarding the variography analysis.
I do not think that Table 1 adds much.	The original Table 1 has been moved to the end of the end of the paper on Page 20 of the revised manuscript.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Authors):

The authors replied in great detail to address my two major comments.
First, they performed an extensive ablation study (section 4.5.2) to understand the impact on generalisation of depth, size of the hidden layers, convex combination and covariate scaling layers. I also found useful the runtime comparisons in Fig. 7.
Second, they clarified the role of the variance of the normal prior over parameters, which I missed in my first report. In this respect, I think it is a wise choice to promote the variance to be a learnable parameter.

Concerning their open-source BayesNF software, I feel that the authors addressed most of my criticism and improved usability.

I also went through the report of Referee 2, who highlighted a substantial number of possible improvements. My overall feeling is that the authors' reply is in most cases valid and scientifically sound, and their new analyses look rather convincing.
I found instructive to read the reply to point 3 in Referee 2 report and the inclusion of a comparison with the NBEATS baseline.

Finally, I must confess that reading the report of Referee 2 encouraged me to go through the literature on spatiotemporal predictions to improve my understanding of the state of the art.
I found this (very) recent manuscript by another research group at Google (https://www.science.org/stoken/author-tokens/ST-1550/full), where the authors propose a novel method for global weather forecasting.
At a superficial reading, I feel that this manuscript is related to the work by Saad et al., and I would be glad if the authors could discuss this in more detail. The two first questions that come to my mind are the following: (i) could the authors highlight the differences between their approach and the one by Lam and co-workers? (ii) Is it possible to make comparisons between the two proposed algorithms?
As far as I understand, the code of the model GraphCast used in the work is available at this link: https://github.com/google-deepmind/graphcast

Reviewer #1 (Remarks on code availability):

The authors definitely improved the quality of the code after the first round of review, implementing all my requests.

Reviewer #2 (Remarks to the Authors):

I believe that the updates for this resubmission are generally excellent and have substantially improved the paper. The ablations in Figure 7 are a particularly nice addition that have convinced me of the efficacy of the specific algorithmic setup chosen. More generally, it is clear the concerns raised in my original review have been taken seriously and the authors have done a very good job of addressing them. With the updates, I am therefore now happy to back acceptance of the paper and do not have any substantial requests for further updates.

Couple of minor points:
- Though the motivation in the intro and at the start of the method introduction has definitely been improved, I think there is still some room for making this clearer still. In particular, I think that more could be done to make it as clear as possible a) the key ways the approach differs from a generic BNN setup, and b) the high-level motivation for these innovations.
- Around lines 537-538 there seems to be a contradiction of whether the ablation is adding/removing a single layer or doubling/halving the number of layers.

Reviewer #2 (Remarks on code availability):

I have not re-reviewed the code in this updated submission, but did check it in my previous review.

We thank the reviewers for their careful consideration of our paper and their comments regarding our revised manuscript. We briefly respond to some final comments from the updated reviews.

Reviewer 1:

> (i) could the authors highlight the differences between their approach and the one by Lam and co-workers? (ii) Is it possible to make comparisons between the two proposed algorithms?

GraphCast (Lam et al., 2023) is a domain-specific model that is developed only for weather prediction problems. In contrast BayesNF is a domain-general modeling tool that is easily applied to a wide spectrum of geostatistical prediction problems in datasets (e.g., weather, pollution, disease, etc.) that follow a given spatio-temporal format (i.e., so-called "long form" panel data) analogously to how practitioners can easily apply Gaussian processes or generalized linear mixed models in these settings.

GraphCast is based on Graph Neural Networks (GNNs): a broad discussion of the differences between GNNs and BayesNF is given in lines 142--152 of the revised submission. Consistent with our observation "GNNs have not yet been demonstrated on probabilistic prediction tasks, and we are unaware of the existence of open-source software libraries based on GNNs that can easily handle the sparse datasets in Section 2.1.", the authors of GraphCast note that uncertainty quantification and formulating probabilistic predictions remains a key limitation of their model, writing:

> "By contrast, GraphCast's MSE training objective encourages it to spatially blur its predictions in the presence of uncertainty, which may not be desirable for some applications where knowing tail, or joint, probabilities of events is important. Building probabilistic forecasts that model uncertainty more explicitly, along the lines of ensemble forecasts, is a crucial next step."

Based on these differences, GraphCast is not directly applicable to our Evaluation, which contains several datasets beyond short-term weather prediction and evaluates probabilistic forecasts based on 95% prediction intervals using the mean interval score (MIS).

Reviewer 2:

> Around lines 537-538 there seems to be a contradiction of whether the ablation is adding/removing a single layer or doubling/halving the number of layers.

These lines have been clarified.