Response to reviewer comments on “Deep learning for stochastic precipitation generation – Deep SPG v1.0” by Bird et al.

Review 1

Reviewer comments are shown in blue and our responses are shown in black.

This manuscript presents a regression-style deep neural network that serves as a single-site stochastic precipitation generator. The neural network relates daily (or hourly) precipitation to previous days (hours) precipitation and sinusoidal terms to capture seasonality. The output of the neural network is a mixture distribution with four components. The manuscript shows that the generator is able to match the statistical characteristics of observed precipitation series well. The manuscript then goes on to explore how to represent non-stationarity from climate change, assessing whether the quantiles of precipitation robustly scale with temperature and concluding that the estimated relationships may not be robust. Finally the manuscript scales the generated time series from the stationary model using the relationship with temperature derived from weather@home simulations as an exploration of a possible methodological approach. I am a statistician with experience with climate statistics and so will concentrate on some of the statistical/machine learning aspects of the work. In general, I thought the work was a useful exploration of one way to use machine learning methods for precipitation simulation. It was interesting to see the use of a regression-based neural network rather than a network specifically designed to reproduce temporal structure such as an LSTM. The simulator seems to do a good job of capturing precipitation variation over time. However, I did have a number of questions about the methodological approach and thought there were some areas of lack of clarity in terms of the methods.

1. The neural network modeling framework is complicated when one considers all the choices made in terms of the network structure (Fig. 1), choice of predictors (Section 3.1), the mixture distribution (a four-component mixture), and the optimization parameters (lines 219-221, 225-228). The authors do not describe at all how any of these choices were made, yet these choices are fundamental to the work and represent a potentially important contribution to the literature. Were these choices made empirically, based on trying different approaches on data (which could lead to overfitting, depending on how this was done)?

We agree with the reviewer that we did not go into sufficient detail on some of the methodological choices made and have now expanded on this in the manuscript. Many of the choices made in designing the architecture of the model were made empirically based on trying different approaches on the data. Minimising the loss function was not the only criterion used for selecting an optimal model - we were also careful to avoid overfitting (checking the validation loss), and, in particular, ensuring that no specific model architecture generated unrealistically large (unbounded) extreme precipitation values.

Do the authors have any idea whether other choices would give similar performance?

We do. A **lot** of time was invested in optimising both the architecture of the neural network underlying the SPG and training the SPG. A large tree of methodological choices was explored, though perhaps not always completely, and not always in a formalised way. The identification of many dead-ends in the exploration of that ‘methodological choices tree’ was not documented in detail in the paper as it would have made the paper excessively long. We have, however, now added a little more material to give some sense as to this process.

We get a little bit of information from the comparison to the regression, but the black box nature of the model choice is troubling and the lack of discussion of how the methods were arrived at feels like a big omission.

We have now added more material following Figure 1 to justify the choice of architecture. Many simpler architectures were tested but were found to perform poorly compared to the architecture that was finally settled on and which is outlined in Figure 1.

On a minor note, I didn't really understand the rationale for the extra fully-connected layer (lines 180-181). In a standard neural network, one would map directly from the dimensionality of the inputs to the desired dimensionality of the first hidden layer. What does the more complicated structure at the top of this network achieve?

The extra full-connected layer is required because the dimensionality of the input is not 256. The additional fully connected layer, inserted between the green input block and the concatenator (+ sign in the block), ensures dimensional compatibility at the concatenation stage within each block. We have added two sentences to this effect in the manuscript and have also modified Figure 1 to ensure that this flow is made more clear.

2.) I'm having some trouble understanding the loss results, summarized in Fig. 2 and Section 3.6.
- How can it be that the validation loss for the neural network is approximately constant over the epochs? (Perhaps the scale of the y-axis makes things hard to see?)

The training of the neural networks converges very quickly such that much of the training is indeed achieved within the very first epoch. While there is a minimum around epoch 5 or 6, the NN behaviour results in a relatively small decrease in the mean negative log-likelihood from the training in the first epoch to the minimum in epochs 5 or 6.

Shouldn't the optimization result in a decrease in validation for the initial epochs?

It does but most of the decrease occurs through the epoch 1 training, i.e. the neural network has mostly converged after a single epoch. This is more so the case for the hourly SPG where there is a large volume of data, even in the first epoch. We have added a sentence to this effect.

Given this, it's hard to understand the statement (line 265) about 10 epochs, as the validation loss for the daily model seems to have a minimum at the first epoch.

No, the minimum is not at the first epoch but rather around epoch 5 or 6 for the daily SPG and at or before epoch 10 for the hourly SPG. This is hard to see in Figure 2 because the neural network has largely converged during the epoch 1 training such that the mean negative log-likelihood shows only very small decreases thereafter to the minimum around epoch 5 or 6 for the daily SPG and epoch 10 for the hourly SPG. For both the daily and the hourly SPG, the mean negative log-likelihood starts to rise out of the noise after about 10 epochs. We have added a sentence and clarified the text in this regard.

I'd also like to understand how the validation was done. Was this basically one-step ahead prediction, using the actual observations as the inputs for each of the 1000 (or 10000) validation observations.

Yes it is one-step-ahead prediction over a year of continuous data, i.e., it is not split randomly. The features are calculated using the validation data (e.g. how much precipitation was over the past 8 days) and then using that to predict the likelihood of the next precipitation occurrence. A sentence has been added to Section 3.4 to clarify this.

- The losses seem to be the average loss per observation.

Yes, this is correct.

If one considers the total loss, one might be able to think more about whether the difference between the linear and neural network models is important. For example if one has nested statistical models, one can use a likelihood ratio test to assess whether the additional parameters in a more complicated model are warranted given how much better the loss is compared to a simpler model.

Yes, we understand, e.g., the application of some metric such as the Bayesian Information Criterion (BIC). However, because we are using hold-out validation, we are assessing these models independently of their training loss and number of parameters.

One can't do that exactly here, but with 1000 observations, I think that the negative log-likelihood for the neural network is 926 compared to 932 for the linear model. That doesn't actually seem to be a very big difference (considering the usual chisquare statistics used in likelihood ratio test), so I'm not fully convinced that (particularly for the daily case) a neural network is really doing that much better than a simpler model.

For both the hourly and daily SPG this isn’t relevant since we haven’t validated using these methods. We agree that for the daily SPG there isn’t a big difference between the NN-based model and the linear model in **terms of loss**. However, this criterion was not used for selection, i.e. the validation loss was used rather. We are not fitting in the same way as for a statistical model where you seek to find the optimal solution. As we say in the paper, there is only a small advantage of using a neural network for the daily case. However, it does still perform better given the results of our hold-out validation, which is a more reliable estimate of performance. For the hourly SPG the neural network does perform considerably better than the linear model.

If the validation data were used for early stopping, then they can't really be used to compare the neural network and regression models. One would presumably need a threeway split of the data, with some test data fully held out of the fitting process, including the early stopping determination.

We agree. However, it is clear that no matter at what point we stop the training, there is a significant difference between the linear model and the neural network (see Figure 2). As such, the additional complexity resulting from the three-way split of the data would be unnecessary. Given the small variability in the validation loss, it is clear that the performance of the neural network does not depend critically on where the training is stopped. Ultimately, the selection of the model was determined primarily by the coverage of the distribution and, in particular, the ability of the model to represent the extremes. We have therefore not added anything to the text in this regard.

3.) The assessment of the simulated series against observations explores various aspects of the pattern of precipitation and is generally (though not entirely) reassuring in terms of the simulations matching the statistics of observed precipitation. However, I think the comparison could be improved by simulation more time series, since the whole point of a generator is the ability to generate as many time series as desired. For example, with many simulated series, the authors could show predictive uncertainty bounds in the quantile-quantile plots (e.g., Fig. 3) to better understand if the divergence between the simulated and observed series at high quantiles might be explained by stochasticity to some degree.

We thank the reviewer for this suggestion. We have now generated a 13,000 year ensemble of SPG output and, from that, calculated a range of statistics that now appear in four new tables that have been added to the manuscript. We believe that the assessment of the extremes displayed in those tables quantifies the spread in the QQ plots at the extremes.

Similarly, the autocorrelation plots show a lot of wiggliness (particularly Fig. 5) because of the limited number of data points. This is unavoidable for the observations, but if one simply simulated more series, one would probably get autocorrelation estimates that vary smoothly with time lag for the generator.

In this case we wanted to keep the SPG results as comparable as possible to the observations and therefore did not want to smooth the SPG results by way of generating a very large ensemble.

4) I found the sections on non-stationarity somewhat hard to follow.

- Line 344 says that within each percentile the data were averaged over each calendar year. I'm not sure what this means. If we take an observed time series and calculate, say, the 60th percentile within each year, that is a single number for each year. So I don't know what numbers are being averaged in each year.

We agree that this sentence was confusing. We have reworded the material at the top of Section 5.1 to add greater clarity and deleted this sentence.

- Related to this, I don't know where the uncertainty bands in Figs. 11-13 come from.

The uncertainties in Figures 11 to 13 were calculated from uncertainties in the fit of the $p(T) = p_0 e^{rT}$ equation to the values within each quantile. A sentence has been added to the top of Section 5.2 in thus regard.

- I generally found Section 5.3 hard to follow. In particular, I can't tell if the text in lines 387-391 is meant to summarize the text that follows, or if it describes prepatory steps that precede the steps described in the steps that follow. I.e., are the 'correlation' and 'scaling' discussed here what is described in more detail in lines 406-409?

These were preparatory steps and provide a prelude to the text that follows. Yes, the 'correlation' and 'scaling' discussed here are what is described in more detail in lines 406-409. We have added a phrase that better connects the two pieces of material describing the scaling.

5) In addition, I had some concerns about the formulation of non-stationarity.

- It would be helpful to see some evidence (e.g., based on scatterplots of input data overlaid with the fit) of whether the functional form relating precipitation to temperature (line 347) fits observed/modeled data well.

Figure 14 does provide this to some extent, though it shows the ensemble mean quantile value rather than the individual ensemble quantile values derived for each year (noting that each ensemble is for a single year) at each T'_{SHland} value. However, we did not want to extend the paper unnecessarily by adding further such plots. We have added additional explanatory text regarding Figure 14.

- Particularly for the observations, could there be other factors (in particular aerosols) that affect precipitation and might be correlated with temperature?

Yes, it is likely that other factors also have an impact on precipitation, however it would be even more difficult to isolate many potential factors and ultimately is another reason why learning the impact of climate change from observations should be avoided, as we show in the paper.

- Are the relationships in Fig. 16 (particularly for Christchurch) believable?

We have no reason to believe these rates are unreasonable, decreases in total precipitation are expected in the far north (Auckland and Tauranga). While increases in the extremes are largely consistent and approach Clausius–Clapeyron scaling. The hook structure for Christchurch while unusual does not seem unreasonable, given there may be complex changes in precipitation dynamics due to climate change.

- The analysis of non-stationarity and the scaling relationships don't seem to account for seasonality. Would the scaling relationships be expected to be the same for different seasons?

We did not account for seasonality in the post hoc approach and did not investigate this issue as we did not expect large differences in scaling. However, this would certainly be something to investigate in the next version of the SPG.

- The discussion highlights the potential of the generator for use in climate change assessment -- this seems to depend critically on being able to estimate the scaling relationship with temperature, which the work casts some doubt on.

Yes, this work casts doubt on the ability of the SPG to learn the climate change scaling from historical data. However, we present a post hoc approach with a scaling derived from a climate model, that when combined with the SPG may still provide utility for climate change impact assessment and down stream modelling.

Reviewer 2

Reviewer comments are shown in blue and our responses are shown in black.

The manuscript by Bird et al. presents a deep learning based stochastically generation of daily and hourly precipitation time series in New Zealand. The Authors used a neural network framework and probabilistic mixture distributions (i.e., gamma and GPD) to simulate the precipitation intensity at daily and hourly timescales. A list of statistics is shown to examine the synthetic time series and model performance against the observations and future climate projection. The Authors also present a non-stationary version of their deep learning model that incorporates contemporary/future changes in the precipitation traces through a temperature covariate. Overall, this manuscript contributes by adapting a neural network approach to stochastically develop multiple ensembles of precipitation at the regional scale. My comments/questions are listed below that are mostly related to the model structure, performance assessment, and precipitation temperature sensitivity parts:

We thank the review for taking the time to review the manuscript.

1) What was the reason for making the location parameter of GPD zero (L200)?

A GPD is typically used for modelling the tails of a distribution, however each component of the mixture model needs to be able to support the full range of potential precipitation depths when calculating the log-likelihood. For this reason, we set the location parameter to zero, despite this somewhat unusual use of a GPD distribution it did improve the representation of the extreme compared to only including Gamma distributions.

Also, were the parameters estimated at a seasonal scale

The model does have a seasonal signal, one of the input features to the NN is a sine and cosine term representing the seasonal signal. The NN learns how the distribution should change with respect to these terms.

or any other specific considerations applied to not mix up different rain-generating mechanisms into the same distribution?

There are no mechanism specific consideration, however the neural network is free to learn different mechanisms if they have an impact on the distribution of precipitation in the observation time series.

I presume a threshold or set of thresholds should also be considered to distinguish between the mixture distributions (lower quantiles vs. heavy tail-like behavior); how did you set the model to learn about these limits?

The distribution that is learned is inherently heavy-tailed so no specific distinction needs to be made. We uniformly sample from this distribution.

2) I was expecting that the developed stochastic precipitation generator model could also produce traces of precipitation reasonably higher or lower the observed time series across different ensemble members.

The SPG can produce values higher than observed in the training data set but not lower since it cannot produce precipitation values lower than 1mm / day for the daily SPG and 0.1 mm/hour for the hourly SPG.

However, it seems any values greater than the observed maximums (L249-252; Table 5) are replaced with equal or smaller magnitudes than the recorded maxima at daily and hourly timescales.

No, these are not observed maximums. These values are considerably higher than the previous events. They are simply upper bounds, which stop the SPG from producing totally unrealistic precipitation depths, which can very occasionally occur when uniformly sampling from the mixture model.

3) I would strongly suggest providing a few quantitative measures for the model comparison (e.g., percent bias, index of agreement, and so on) in addition to showing the model performance q-q plots in the “Stationary quality assessment” section for different precipitation characteristics, including moments, spells, extreme values of simulated traces. It is unclear how the stochastic model (e.g., different ensemble members) performs against the observations by only looking at these q-q plots.

We have now included metrics for dry days, moments, and extreme values comparing daily and hourly precipitation between observations and SPG simulations (Tables 6-9). We did not include an index of agreement as it would not be appropriate as the SPG doesn't provide a time series that is aligned with the observations. Furthermore, we thought including the metrics for the observations and the SPG is more informative than the percent bias.

4) How does the stochastic model preserve the year-to-year variability of seasonal/annual precipitation variation?

We have added the standard deviation of annual mean precipitation, for both the daily SPG and for observations, as a diagnostic in Table 8 for each of the four sites. Interestingly, there is no significant difference in the standard deviation in the annual mean precipitation derived from the dialy SPG and from the daily observations. This suggests that forced (as opposed to random) inter-annual variability in precipitation is rather small across these four sites in New Zealand - although an in-depth investigation of this conclusion is beyond the scope of this paper. For the hourly SPG, on the other hand, the variability in annual mean precipitation is underestimated at all locations. We have added some discussion of the annual variability to Section 4.5.

Can the Authors compare the intra-annual variation between the simulated traces and observed time series?

We have added the standard deviation of annual mean precipitation, for both the daily SPG and for observations, as a diagnostic in Table 8 for each of the four sites. Interestingly, there is no significant difference in the standard deviation in the annual mean precipitation derived from the SPG and from observations. This suggests that forced (as opposed to random) inter-annual variability in precipitation is rather small across these four sites in New Zealand - although an in-depth investigation of this conclusion is beyond the scope of this paper. For the hourly SPG, on the other hand, the variability in annual mean precipitation is underestimated at all locations. We have added some discussion of the annual variability to Section 4.5.

5) precipitation-temperature sensitivity: When I look at the P-T sensitivity plots, it seems negative scaling rates are calculated in many stations.

Correct.

Note that the Clausius–Clapeyron equation demonstrates a ‘positive’ relation between an increase in temperature and rainfall when the atmosphere is (nearly) saturated.

Yes, we are aware of this ,i.e., the Clausius–Clapeyron relationship suggests a +7%/K increase in the water carrying capacity of the atmosphere. We note, however, that this is different to a +7% increase in precipitation in that it is not only the water carrying capacity of the atmosphere that determines precipitation, but also climate-induced changes in the dynamics that may affect **where** the rain falls - in reality it is not unusual in many places for climate warming to induce decreases in light precipitation and increases in heavy precipitation.

As an example, it seems there is a strong ‘hook’ structure in the P-T relation across your stochastic precipitation generator/observation/model precipitation time series, and it should be taken into account before calculating the rates.

We agree and the hook was taken into account when calculating the rates, i.e., the hook structure is imposed onto the SPG.

6) “5.3 Post hoc addition of non-stationarity”: It is unclear whether the established P-T relationship is valid given Figures 16-17 and Comment#5 above.

We believe the figures are valid, given that the hook structure observed in Figure 16 was imposed onto the SPG when it was used to generate Figure 17.

Additional comments:

L117-124: I doubt using a single weather@home grid cell is a robust approach here in general; how about using at least the four nearest grid cells to the site/station? This way, the failure rate of encountering a ‘bad’ cell will statistically decrease fourfold.

We have used weather@home for several studies and, indeed, often use the average of the four nearest grid cells to the location of the observations as representative of the precipitation. However, for this study we only use the weather@home data to infer the sensitivity of the precipitation to a climate covariate, in this case the Southern Hemisphere land temperature anomaly - a field expected to be spatially more homogenous than the precipitation field itself.

Furthermore, it wasn’t quite clear to us what the reviewer meant by a ‘bad’ cell. In our analyses of the weather@home data, we do sometimes find a cell with an occasional unrealistic precipitation value. However, we did screen the weather@home data to avoid such outliers. Our derivation of the sensitivity of the precipitation to such ‘bad’ cells is therefore mitigated.

L183-186: What computational settings were used to execute this neural network and generate the traces? Please add some details about the logistics/computing configuration requirements.

The model was trained on a 24-core threadripper CPU. The model only took a few minutes to train on this machine. It could also be trained on a GPU. If the reviewer is asking about the python environment used for training the SPG, we have provided the conda environment .yaml file in the repository pointed to in the manuscript.

L196-203: Were these 12 parameters estimated for each site without considering any seasonal influence on precipitation distributions? e.g., cold vs. warm season.

Yes, one of the input features to the neural network is a sine and cosine term representing the seasonal signal. The neural network learns how these parameters should change with respect to these terms.

L221-222: It is unclear what “better numerical stability” means here.

We have changed the wording, optimization using the likelihood requires calculating a joint product which quickly leads to floating point underflow for small probabilities, instead we optimize using log-likelihood which instead requires a sum, avoiding floating point underflow.

L234-235: Is it causing a problem if the first eight days or 144 hours happen to be all Zeros?

No. We anyway often observe dry period spells longer than 8 days and the model still exits those dry spells, creating dry-spell lengths consistent with the distribution of dry spells seen in reality (see Figure 7).

L312: What is your hypothesis on the hourly SPG that underestimates the seasonality in the proportion of dry hours for Auckland?

We have no hypothesis but acknowledge that this underestimation could reduce with retraining of the SPG.

L321: Is this the entire section of the “4.5. Discussion” following the “4. Stationary quality assessment”? I feel this “4.5” section does not add any new information more than what is presented in “4.1” to “4.4” sections.

This is a good point made by the reviewer. We have deleted this section and ensured that any material covered by it now appears elsewhere.

In general, it seems some figures can be merged, or removed, as they do not add any new information.

We carefully assessed and considered each figure and couldn’t identify any that are surplus to the requirements of the paper.