Convolutional conditional neural processes for local climate downscaling

Author response

Referee #1

The paper applies a convolutional conditional neural process (convCNP) to the task of statistical downscaling, in particular of temperature and precipitation.

Methodology falls within the realm of probabilistic deep learning, allowing to combine a bespoke statistical model with deep learning framework.

Extensive experiments are performed to compare the convCNP model to a range of benchmarks, with promising results.

While there are a number of modelling choices and experimental details in this paper that would benefit from a much better motivation and explanation, this is undoubtedly a good contribution to the literature.

28-29 "This is based on the assumption that while sub-grid-scale and parameterised processes are poorly represented in GCMs, the large scale flow is generally better resolved."

-- Can this assumption be elaborated further or an appropriate reference included?

An appropriate reference has been added (Maraun and Widmann, 2018).

83-84 "This is assumed to be Gaussian for maximum temperature and a Gamma-Bernoulli mixture for precipitation"

These choices are made for consistency with existing stochastic downscaling models in the literature. This was indeed unclear in the original manuscript, we have added a line explaining this in Section 2.1 with appropriate references (Gutierrez et al., 2013; San Martin et al., 2017; Cannon et al., 2008).

99 should say "distributional parameters θ at each target location"

Fixed

118 "parameterise a stochastic process over the output variable, in this case either temperature or precipitation"

-- Can it be made explicit how exactly does the model proposed here result in a stochastic process over the output variable since this is obscured by a large number of model components, and if there is anything in the proposed model that is in fact "nonparametric"? The overall model reads as a complicated, but a parametric model for the parameter of the conditional distribution. I found this paragraph confusing since it appears to contrast the

proposed method to "parametric approaches". Also comment on what are the advantages of having such a stochastic process in this specific context?

Thank you for this comment, this was indeed unclear in the original manuscript. We were referring to parametric in a machine learning sense i.e the transfer function is parameterised using a neural network. This was not meant to suggest that the model for temperature and precipitation is nonparametric. We have rewritten Section 2.1 to give a clearer description of the convCNP model.

154-155 "The VALUE experiment protocol does not specify which predictors are used in the downscaling model (i.e which gridded variables are included in Z)."

-- Can you clarify if all the baselines in comparisons use the same set of predictors? Do any of them use topographic predictors? If none of them use topographic predictors, would it be more fair to compare convCNP without topographic predictors to the baselines, and then consider the (additional) improvement due to including topographic information?

Thank you for this comment. All the baselines use different predictor sets, we have included a reference for this. In the VALUE experiment a separate model is trained for each location, hence topography is not included in the predictors as it remains constant. We therefore believe that the experimental setup is fair. We have added the text

It is emphasised that in the VALUE baselines a separate model is trained for every location, hence topographic predictors are not required.

in Section 2.2.

Fig.10 Plotting KDEs is highly problematic here due to bounded domain. It would be much more informative to simply plot a histogram of the evaluations of CDF at the true values. Perhaps perform a KS test to check if distribution is uniform in "well calibrated" cases?

Thank you for pointing this out. After consideration of both referees comments we opted to replace the KDE plots with histograms, and use a simple visualisation as opposed to statistical tests for uniformity.

Remarks on notation:

Thank you for spotting these typos, we have made the suggested corrections in the revised manuscript.

References

1. Maraun, D. and Widmann, M.: Statistical downscaling and bias correction for climate research, Cambridge University Press, 2018.
2. Gutiérrez, J. M., San-Martín, D., Brands, S., Manzanas, R., and Herrera, S.: Reassessing statistical downscaling techniques for their robust application under climate change conditions, Journal of Climate, 26, 171–188, 2013.
3. San-Martín, D., Manzanas, R., Brands, S., Herrera, S., and Gutiérrez, J. M.: Reassessing model uncertainty for regional projections of precipitation with an ensemble of statistical downscaling methods, Journal of Climate, 30, 203–223, 2.

4. Cannon, A. J.: Probabilistic multisite precipitation downscaling by an expanded Bernoulli–Gamma density network, Journal of Hydromete- orology, 9, 1284–1300, 2008.

Referee #2

This paper presents a novel method for statistical downscaling leveraging the recent literature on neural processes, or more specifically, convolutional conditional neural processes (ConvCNPs). The authors present comprehensive experimental results for downscaling temperature and precipitation based on the well-established VALUE downscaling intercomparison framework.

Neural processes were undoubtedly a significant step forward in the field of probabilistic deep learning, and their value for the task of statistical downscaling is quite clear. Thus, the authors' work absolutely represents an important contribution to the literature and provides ample motivation for the continued investigation of ConvCNPs for downscaling, as well as related tasks in the future. The paper is also generally well written and the results are clearly presented.

However, despite its strengths, I do not think that this work is ready for publication as submitted. The literature review is insufficient and omits very relevant recent works in the field, e.g [2,4,7,8,9]. As a result, the authors significantly overstate their contributions on several occasions, in particular with regards to multi-site downscaling and generalization. Even worse, they misrepresent the current state of research on applications of deep learning in downscaling as somehow pessimistic, drawing a questionable contrast between the reported success of their method and the results of prior work. In fact, numerous authors have had success in applying deep learning to statistical downscaling tasks in recent years (see other references); thus the statement "...previous work [suggests] that little benefit is derived from applying neural network models to downscaling…" is, at best, misleading in the broader context of the literature.

Thank you very much for pointing out these papers. It is much appreciated and addition of these references has significantly improved the literature review. Whilst we agree that the initial draft of the manuscript was too pessimistic on the state of downscaling research using deep learning, we do note that several recent works have questioned whether deep learning delivers significant improvements over simple statistical techniques for temperature (Bano-Medina, 2020) and precipitation (Vandal 2019), and works published since we submitted our paper have expressed similar sentiments in their introductions. For example Wang et al. (2021) state that "*While the downscaling results generally indicate improvement, the performance of the plain CNN models were not consistently better than the classic statistical techniques*" in their review of the current state of the literature. We have updated Sections 1 and 6 with the papers you suggest to present a more balanced view.

Similarly, the repeated claim that "existing downscaling methods are unable to handle unseen locations" is not strictly true. It is typically possible, though not always effective, to train a downscaling model on one location and test it on another, provided that similar low resolution predictors are available. As noted in [1], convolutional neural networks are already

well equipped for this when trained with data from multiple sites since they tend to learn more spatially invariant features. Transfer learning is also possible with other deep learning methods, as very recently demonstrated by [10]. ConvCNP may still generally be better, but the experiments presented by the authors do not show this. They only show ConvCNP is superior in comparison to a constructed Gaussian process interpolation baseline.

Thanks for this comment. The convCNP method differs from transfer learning in that the output of the model is a stochastic process (i.e a distribution over continuous spatial functions of the downscaled variable). This can be queried at any set of locations at inference time. In contrast, transfer learning models such as those outlined by Wang et al. (2021) can be applied in new locations, however will always output predictions on a grid/set of points fixed by the training data (we note that the work of Wang et al. (2021) was not available at the time we submitted this paper). The convCNP therefore provides a more flexible approach, which was poorly explained in the original manuscript. We have altered the introduction to emphasize that the novelty of our approach is that the output prediction is spatially continuous, and added a line comparing the convCNP method to transfer learning. We have also added references for other applications of deep learning models which output continuous spatial fields that have successfully been applied to idealised problems (Li et al., 2020a,b; Lu et al., 2019).

This brings me to the last major issue. The experimental analysis, while otherwise fairly rigorous and well designed, has one significant weakness: it does not, as far as I can tell, compare the authors' proposed ConvCNP method to any other deep learning methods recently proposed in the literature, including those that they cite such as the CNN architecture proposed by Baño-Medina et al. While this is not necessarily an absolute requirement for this paper to be a valuable contribution, it is at odds with how the authors seem to want to place their work in the context of current research. The authors suggest in sections 1, 2, and 6 that their approach is superior to existing deep learning methods. While this indeed may be true, it is not supported by their current results. I suggest that the authors should either add one or more existing deep-downscaling methods to their experiments, or re-frame their work and reduce the scope of their claims.

Thank you for your feedback on this point. As the convCNP operates on ungridded data (i.e raw station observations) direct comparison with convolutional architectures such as those presented by Bano-Medina (2020) and Vandal (2018) is challenging as these models require input on regular grids. We therefore opted to use traditional statistical baselines from the VALUE experiment (where a separate model is trained for each station), and construct an interpolation baseline using a Gaussian process as this is a commonly used baseline for neural process models across a range of tasks (Garnelo et al., 2018a, Garnelo et al. 2018b, Gordon et al., 2019). As these simple statistical methods are widely used in climate impact studies we believe that this set of baselines is suitable for the current study, with the caveat that further work is required to develop a standardised framework to compare machine learning models for downscaling and assess the performance of the convCNP compared to these methods.

As a direct comparison is not possible, we agree with your feedback that it is best to reframe this work as presenting a new downscaling model emphasizing the novelty of generating

continuous spatial predictions as opposed to motivating applications of deep learning in downscaling. We have removed the words

"*It is instructive to consider these results more broadly in the context of the debate surrounding the application of deep learning models in downscaling. In this study we deploy deep learning approaches using domain-specific knowledge to shape the architecture, and then deploy this inside simple statistical observation models. This approach is demonstrated to improve on simple baselines on both mean and extreme metrics for maximum temperature and precipitation. This contrasts to previous work suggesting that little benefit is derived from applying neural network models to downscaling maximum temperature (Bano-Medina, 2020) and precipitation (Vandal, 2019), and motivates continued efforts to develop and benchmark such models for application in climate impact studies.*"

From section 6, and instead replace this with a discussion of future work we are conducting to evaluate performance of deep learning models for downscaling and ascertain how the convCNP and other neural process models perform.

In summary, while the authors' work is undoubtedly a valuable contribution, the current presentation and framing within the context of the literature has significant problems. There is also a lack of much needed detail in the description of the ConvCNP model. The paper would benefit from a description which highlights step-by-step the similarities and differences of their architecture with the original ConvCNP architecture described by Gordon et al, as well as with other similar architectures like the ones described by Baño-Medina et al.

I will summarize my remaining technical comments by section. I look forward to seeing the authors' comments and revisions.

Thank you for your extensive comments, we address each of these separately below:

Section 1

- Similar to what was mentioned previously, the statement "there has been debate as to whether deep learning methods provide improvement over traditional statistical techniques such as multiple linear regression" is perhaps overly pessimistic. There has been plenty of work now that clearly establishes the value of deep learning in statistical downscaling. Citing just two papers which happen to show somewhat mixed results in some cases comes off as a bit disingenuous to anyone who is familiar with the literature.

Thanks for this feedback, we have addressed this point above.

- "The second limitation common to existing downscaling models is that predictions can only be made at sites for which training data are available." Again, this is just simply not true. Nothing stops you from applying a model trained on location to another location where low-resolution predictors are available. How well it generalizes depends on the type of model used (e.g. a dense neural network will probably overfit) and geographic similarity. Where input or output grid interpolation is necessary, it is true that this can introduce additional error, but it is still certainly possible to generate "off grid" predictions.

As discussed above, we have added a discussion of transfer learning to the introduction and explained how this differs from the convCNP model. The advantage over CNN based methods is that we can train directly on station data without introducing errors via regridding, and generate point predictions without requiring interpolation at test time.

- It would be helpful to provide precise definitions for “on-the-grid” and “off-the-grid” in the context of downscaling.

For clarity, we have opted to replace references to on-the-grid and off-the-grid with regularly sampled and irregularly sampled spatial locations.

Section 2

In response to your comments below we have decided to restructure section 2.1 to provide an overview of neural process models prior to explaining the downscaling architecture in detail, which we feel improves the clarity of the section.

- “ψ_MLP is a multi-layer perceptron, φ_c is a function parameterised as a neural network and CNN is a convolutional neural network.” This sentence is a bit confusing. MLPs and CNNs are both types of neural networks. So you should be similarly more specific with what kind of neural network parameterizes φ_c.

This was actually a typo, phi_c is a kernel function with learnable length scale parameter. We have fixed this in the revised manuscript.

- The definition of φ_c given in step 2 does not appear to be a neural network but rather a standard squared exponential kernel. The h term is produced by the CNN, but this was discussed separately. You should clarify if and where an additional neural network is used here (and what the optimized parameters are).

Yes that is correct, phi_c is an exponential kernel to guarantee translation equivariance, we apologise that this was unclear due to the typo mentioned above. No additional neural network is used in this step.

- The EQ summations are missing upper bounds. There must be some finite limit to the values of m,n computed here. Probably it is bounded by the size of the input (reanalysis) grid, but this should be stated explicitly as it has very significant implications on runtime complexity.

Thank you for spotting this, the upper bounds are indeed the size of the input grid and have been added to the equation.

- It's not immediately clear why the EQ kernel is justified in this context. Is geographic proximity really that reliable a marker of similarity? Very distant regions can be similar and neighboring regions can be very different. Perhaps such distinctions are learned implicitly by the encoder. But regardless, a brief discussion of this question would be informative.

Use of the EQ kernel ensures that the predictions are translation equivariant, we have noted this in Section 2.1.

- “parameters at each target location θ” → “parameters, θ, at each target location"

Fixed

- “Formally, PP downscaling is an instance of a supervised learning problem to learn the function f in equation 1. Approaches to learning such a function have traditionally been split into two categories: …” The following statements make it sound like PP downscaling methods are always either neural networks or Bayesian, which is not true and probably not what you meant. Perhaps clarify whether or not you are talking about deep learning methods specifically (though, this would also not be true, as Bayesian DL methods have also been proposed [9]).

We agree that this paragraph was unclear in the original manuscript. Here we were aiming to provide a brief background to the convCNP model, and discussing general classes of supervised learning models as opposed to models specifically for downscaling. We have restructured Section 2.1 to begin with an overview of the convCNP model followed by a detailed description of the architecture.

- In section 2.1.1, it’s not explicit how the temporal dimension is handled. Is a separate distribution generated for each time step at each location? Or does each location get just one distribution from which samples are taken for every time step? Maybe specify a time index in the notation (at least initially).

Thanks for this comment, a separate distribution is indeed generated for each timestep, we have specified this is section 2.1.1.

- Maybe I missed it, but I don’t think the ConvCNP and NP papers make any mention of using non-Guassian likelihoods like the Bernoulli-Gamma used here. It might be worth highlighting where this fits into the theoretical framework.

Thank you for spotting this, indeed existing literature uses a Gaussian distribution for the conditional neural process. We have added the following sentence to Section 2.1

We note that this is an extension of existing conditional and convolutional conditional neural process models where the predictive distribution is assumed to be Gaussian (Garnelo et al, 2018; Gordon et al., 2019).

Sections 3 and 4

- Add units to the captions or y-axis of plots.

Fixed

- MAE and bias tend to be tricky metrics for precipitation due to the heavy tails of the distribution and the prevalence of zero values on dry days. How is this handled here? Are they only calculated on days with precipitation?

We agree that these metrics are somewhat problematic. Here we selected MAE and bias for consistency with what metrics are reported for the baseline VALUE ensemble (Maraun et al., 2015), and have also been applied in other studies applying machine learning to precipitation downscaling (e.g Vandal et al., 2017). As in the VALUE data, these metrics are calculated over all days. We have added a sentence to the future work section in Section 6 indicating that assessment of precipitation results using further metrics would be beneficial.

- As mentioned by the other referee, KDE plots are inappropriate for the PIT in figure 10, as can be seen by non-flat appearance of the uniform distribution. Q-Q or P-P plots, or even just ROC curves, would be more illustrative. See [9] for examples of evaluating calibration on downscaling precipitation. I would advise against the idea of using a statistical test for uniformity on the grounds that statistical tests are generally somewhat uninformative and typically rely on arbitrary thresholds and questionable assumptions. A simple visualization is sufficient.

Thank you for pointing this out. After consideration of both referees comments we opted to replace the KDE plots with histograms, and use a simple visualisation as opposed to statistical tests for uniformity.

References

1. Baño-Medina, J. L., García Manzanas, R., Gutiérrez Llorente, J. M., et al.: Configuration and intercomparison of deep learning neural models for statistical downscaling, 2020.
2. Vandal, T., Kodra, E., and Ganguly, A. R.: Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation, Theoretical and Applied Climatology, 137, 557–570, 2019.
3. Wang, F., Tian, D., Lowe, L., Kalin, L., and Lehrter, J.: Deep Learning for Daily Precipitation and Temperature Downscaling, Water Resources Research, 57, e2020WR029 308, 2021.
4. Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A.: Neural operator: Graph kernel network for partial differential equations, arXiv preprint arXiv:2003.03485, 2020a.
5. Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A.: Fourier neural operator for parametric partial differential equations, arXiv preprint arXiv:2010.08895, 2020b.
6. Lu, L., Jin, P., and Karniadakis, G. E.: Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators, arXiv preprint arXiv:1910.03193, 2019.
7. Maraun, D., Widmann, M., Gutiérrez, J. M., Kotlarski, S., Chandler, R. E., Hertig, E., Wibig, J., Huth, R., and Wilcke, R. A.: VALUE: A framework to validate downscaling approaches for climate change studies, Earth's Future, 3, 1–14, 2015.
8. Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., and Ganguly, A. R.: Deepsd: Generating high resolution climate change projections through single image super-resolution, in: Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining, pp. 1663–1672, 2017.

9. Garnelo, M., Rosenbaum, D., Maddison, C. J., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y. W., Rezende, D. J., and Eslami, S.: Conditional neural processes, arXiv preprint arXiv:1807.01613, 2018
10. Gordon, J., Bruinsma, W. P., Foong, A. Y., Requeima, J., Dubois, Y., and Turner, R. E.: Convolutional conditional neural processes, arXiv preprint arXiv:1910.13556, 2019.