Convolutional conditional neural processes for local climate downscaling

Anna Vaughan[1], Will Tebbutt[1], J. Scott Hosking[2,3], and Richard E. Turner[1]

[1]University of Cambridge, Cambridge, UK
[2]British Antarctic Survey, Cambridge, UK
[3]The Alan Turing Institute, UK

Correspondence: Anna Vaughan (av555@cam.ac.uk)

Abstract.

A new model is presented for multisite statistical downscaling of temperature and precipitation using convolutional conditional neural processes (convCNPs). ConvCNPs are a recently developed class of models that allow deep learning techniques to be applied to off-the-grid spatio-temporal data. This model has a substantial advantage over existing downscaling methods

5 in that the trained model can be used to generate multisite predictions at an arbitrary set of locations, regardless of the availability of training data. The convCNP model is shown to outperform an ensemble of existing downscaling techniques over Europe for both temperature and precipitation taken from the VALUE intercomparison project. The model also outperforms an approach that uses Gaussian processes to interpolate single-site downscaling models at unseen locations. Importantly, substantial improvement is seen in the representation of extreme precipitation events. These results indicate that the convCNP is a

10 robust downscaling model suitable for generating localised projections for use in climate impact studies, and motivates further research into applications of deep learning techniques in statistical downscaling.

1 Introduction

Statistical downscaling methods are vital tools in translating global and regional climate model output to actionable guidance for climate impact studies. General circulation models (GCMs) and regional climate models (RCMs) are used to provide

15 projections of future climate scenarios, however coarse resolution and systematic biases result in unrealistic behaviour, particularly for extreme events (Allen et al., 2016; Maraun et al., 2017). In recognition of these limitations, downscaling is routinely performed to correct raw GCM and RCM outputs. This is achieved either by dynamical downscaling, running a nested high-resolution simulation, or statistical methods. Comparisons of statistical and dynamical downscaling suggest that neither group of methods is clearly superior (Ayar et al., 2016; Casanueva et al., 2016), however in practice computationally cheaper statisti-

20 cal methods are widely used.

Major classes of statistical downscaling methods are model output statistics (MOS) and perfect prognosis (PP; Maraun et al., 2010). MOS methods explicitly adjust the simulated distribution of a given variable to the observed distribution, using variations of quantile mapping (Teutschbein and Seibert, 2012; Piani et al., 2010; Cannon et al., 2020). Though these methods are widely applied in impact studies, they struggle to downscale extreme values and artificially alter trends (Maraun, 2013;

25 Maraun et al., 2017). In contrast, in PP downscaling, the aim is to learn a transfer function f such that

$$\hat{y} = f(x, Z) \tag{1}$$

Where $\hat{y}$ is the downscaled prediction of a given climate variable whose true value is y at location x and Z is a set of predictors from the climate model (Maraun and Widmann, 2018). 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.

30 Multiple different models have been trialled for parameterising f. Traditional statistical methods used for this purpose include multiple linear regression (Gutiérrez et al., 2013; Hertig and Jacobeit, 2013), generalised linear models (San-Martín et al., 2017) and analog techniques (Hatfield and Prueger, 2015; Ayar et al., 2016). More recently, there has been considerable interest in applying advances in machine learning to this problem, including relevance vector machines (Ghosh and Mujumdar, 2008), artificial neural networks (Sachindra et al., 2018), autoencoders (Vandal et al., 2019), recurrent neural networks (Bhardwaj

35 et al., 2018) and convolutional neural networks (Baño-Medina et al., 2020; Höhlein et al., 2020). There has been debate as to whether deep learning methods provide improvement over traditional statistical techniques such as multiple linear regression. Vandal et al. (2019) found that machine learning approaches offered little improvement over traditional methods in downscaling precipitation. In contrast Baño-Medina et al. (2020) compared downscaling performance of convolutional neural networks (CNNs) and simple linear models, finding that CNNs improved predictions of precipitation, but did not result in improvements

40 for temperature.

Limitations remain in these models. In many climate applications it is desirable to make projections that are both (i) consistent over multiple locations and (ii) specific to an arbitrary locality. The problem of multi-site downscaling has been widely studied, with two classes of approaches emerging. Traditional methods take analogues or principal components of the coarse-resolution field as predictors. The spatial dependence is then explicitly modelled for a given set of sites, using observations at

45 those locations to train the model (Maraun and Widmann, 2018; Cannon, 2008; Bevacqua et al., 2017; Mehrotra and Sharma, 2005). More recent work has sought to leverage avances in machine learning, for example CNNs, for feature extraction (Bhardwaj et al., 2018; Baño-Medina et al., 2020; Höhlein et al., 2020). Each of these classes of methods has their drawbacks. Traditional methods have limited feature selection, however are able to train directly on true observations. In contrast, deep learning techniques such as CNNs allow for sophisticated feature extraction, however are only able to be applied to gridded

50 datasets. Gridding of observations naturally introduces error, especially for areas with complex topography and highly stochastic variables such as precipitation (King et al., 2013). The second limitation common to existing downscaling models is that predictions can only be made at sites for which training data are available. Creating a projection at an arbitrary location is achieved either through interpolation of model predictions or taking the closest location.

In this study, we address these challenges by developing a new statistical model for downscaling temperature and precipita-

55 tion at an arbitrary set of sites. This is achieved using convolutional conditional neural processes, state of the art probabilistic machine learning methods combining ideas from Gaussian Processes (GPs) and deep neural networks. The model combines the advantages of existing classes of multi-site approaches, with feature extraction using a convolutional neural network to-

gether with training on off-the-grid data. In contrast to existing methods, the trained model can be used to make coherent local projections at any site regardless of the availability of training observations.

60 The specific aims of this study are as follows:

1. Develop a new statistical model for downscaling GCM output capable of training on off-grid data, making predictions at unseen locations and utilising sub-grid-scale topographic information to refine local predictions.

2. Compare the performance of the statistical model to existing strong baselines.

3. Compare the performance of the statistical model at unseen locations to existing interpolation methods.

65 4. Quantify the impact of including sub-grid-scale topography on model predictions.

Section 2 outlines the development of the downscaling model, and presents the experimental setup used to address aims 2-4. Section 3 compares the performance of the statistical model to an ensemble of baselines. Sections 4 and 5 explore model performance at unseen locations and the impact of including local topographic data. Finally, Section 6 presents a discussion of these results and suggestions for further applications.

70 ## 2 Datasets and methodology

We first outline the development of the statistical downscaling model, followed by a description of three validation experiments.

2.1 The downscaling model

Our aim is to approximate the function f in equation 1 to predict the value of a downscaled climate variable y at locations $\mathbf{x}$ given a set of coarse-scale predictors Z. In order to take the local topography into account, we assume that this function also

75 depends on the local topography at each target point, denoted $\mathbf{e}$, i.e

$$\hat{y} = f(\mathbf{x}, Z, \mathbf{e})$$

We take a probabilistic approach to specifying f where we include a noise model, so that

$$p(y|\mathbf{x}, Z, \mathbf{e}) = p(y|\theta(\mathbf{x}, Z, \mathbf{e}))$$

Deterministic predictions are made from this by using, for example, the predictive mean

80 $$\hat{y} = \int y p(y|\mathbf{x}, Z, \mathbf{e}) dy$$

In this model θ is parameterised as

$$\theta(\mathbf{x}, Z, \mathbf{e}) = \psi_{MLP}[\phi_c(h = CNN(Z), \mathbf{x}), \mathbf{e}]$$
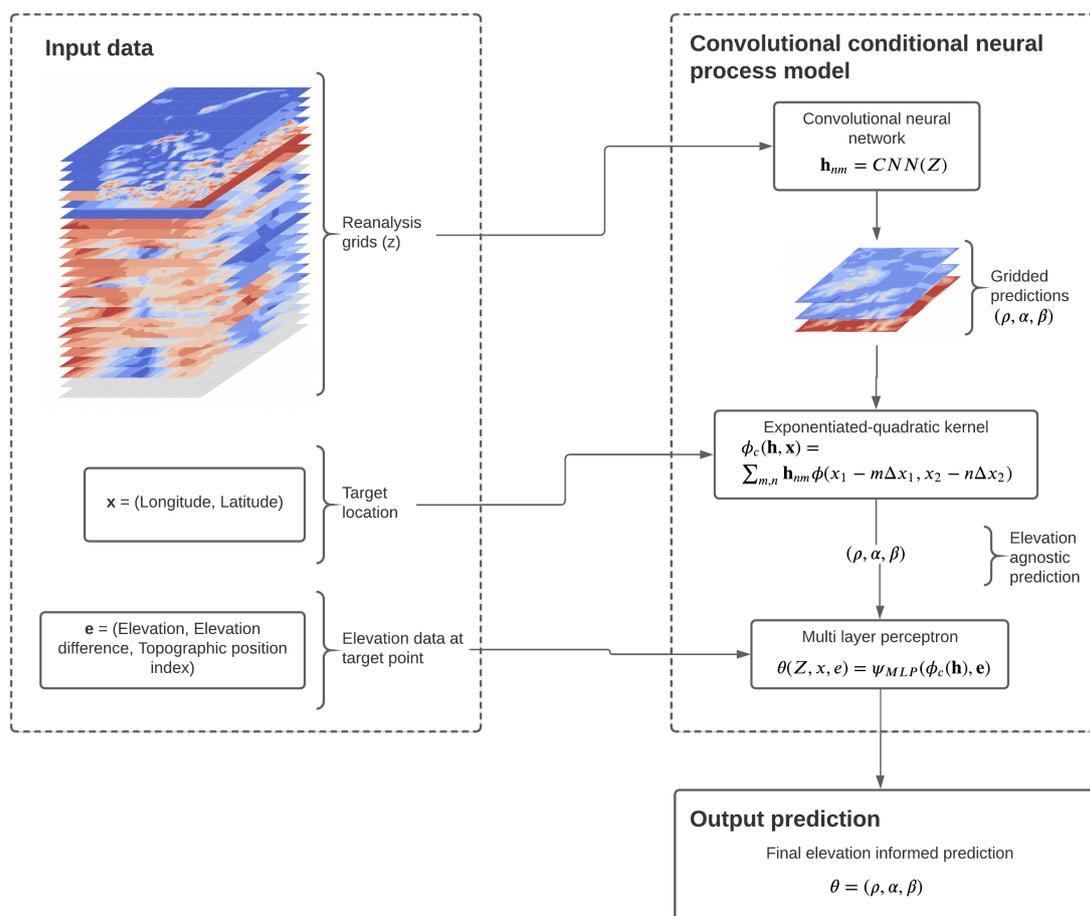
Figure 1. Schematic of the convCNP model for downscaling precipitation demonstrating the flow of data in predicting precipitation for a given day at target locations **x**. Gridded coarse resolution data for each predictor is fed into the CNN, producing predictions of $\theta = (\rho, \alpha, \beta)$ at each grid point. These gridded predictions are then transformed to a prediction at the target location using an exponentiated-quadratic kernel. Finally, these elevation agnostic predictions are fed into a multi-layer perceptron together with topographic data **e** to produce a final prediction of the parameters.

Here θ is a vector of parameters of an distribution for the climate variable at prediction locations **x**. This is assumed to be Gaussian for maximum temperature and a Gamma-Bernoulli mixture for precipitation. **e** is a vector of sub-grid-scale topographic information at each of the prediction locations, ψ_{MLP} is a multi-layer perceptron, ϕ_c is a function parameterised as a neural network and CNN is a convolutional neural network. Each component of this is described below, with a schematic of the model shown in Figure 1.

1. **Convolutional neural network** In the first step, gridded reanalysis predictor data Z are fed into the model. These grids are used as input to a convolutional neural network to extract relevant features. This is implemented as a 6-block Resnet architecture (He et al., 2016) with depthwise separable convolutions (Chollet, 2017). The output from this step is a prediction of the relevant parameters for each variable at each grid point in the predictor set, i.e

$$\mathbf{h}_{nm} = CNN(Z)$$

Where $\mathbf{h}_{nm}$ is the vector-valued output at latitude $m\Delta x_1$ and longitude $n\Delta x_2$ and $m, n \in \mathbb{Z}^+$ with Δx_i indicating the grid spacing.

2. **Translation to off-the-grid predictions** These gridded predictions are translated to the off-the-grid target locations $\mathbf{x}$ using outputs from step 1 as weights for an exponentiated-quadratic (EQ) kernel ϕ, i.e

$$\phi_c(\mathbf{h}, \mathbf{x}) = \sum_{m,n} \mathbf{h}_{nm} \phi(x_1 - m\Delta x_1, x_2 - n\Delta x_2) = \sum_{m,n} \mathbf{h}_{nm} e^{-\frac{1}{2l_1^2}(x_1 - m\Delta x_1)^2 - \frac{1}{2l_2^2}(x_2 - n\Delta x_2)^2}$$

This outputs predictions of the relevant distributional parameters at each target location θ.

3. **Inclusion of sub-grid scale topography**

   By design, the predictions from the previous step only model variation on the scale of the context grid spacing. This elevation agnostic output is post-processed using a multi-layer perceptron (MLP). This takes the parameter predictions from the EQ kernel as input together with a vector of topographic data $\mathbf{e}$ at each target location.

$$\theta(\mathbf{x}, Z, \mathbf{e}) = \psi_{MLP}(\phi_c(\mathbf{h}), \mathbf{e})$$

   The vector $\mathbf{e}$ consists of three measurements at each target point:

   (a) True elevation

   (b) Difference between the true and grid-scale elevation.

   (c) Multi-scale topographic position index (mTPI), measuring the topographic prominence of the location, (i.e quantifying whether the point is in a valley or on a ridge).

   This MLP outputs the final prediction of the distributional parameters θ.

Figure 2 shows a concrete example of temperature and precipitation time-series produced using this model by sampling from the output distributions. Maximum temperature is shown for Helgoland, Germany, and precipitation for Madrid, Spain. For both variables the model produces qualitatively realistic time-series.

This model is an example of a more general class of models known as convolutional conditional neural processes (ConvC-NPs; Gordon et al., 2019). 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: parametric approaches,
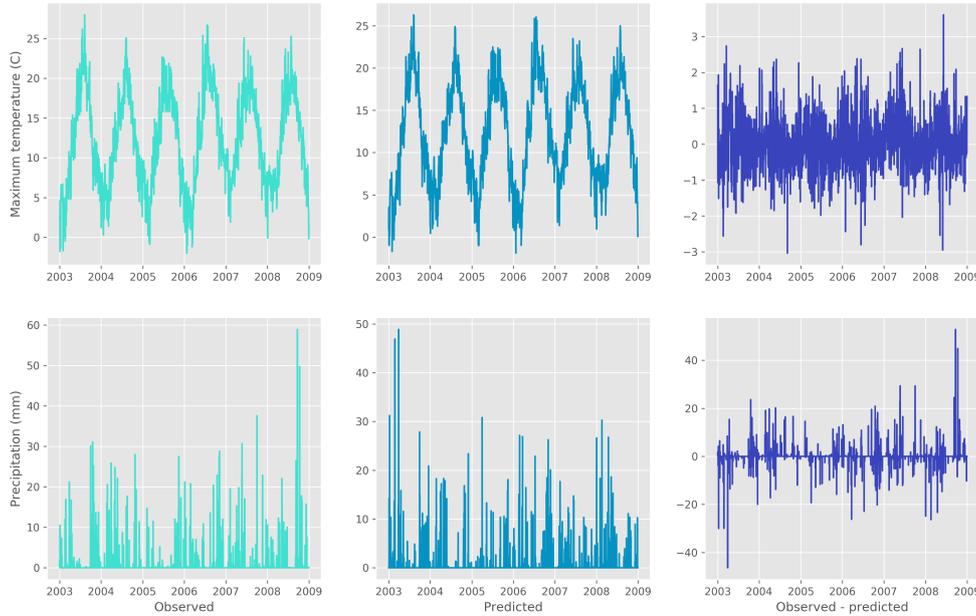
Figure 2. Examples of convCNP model predictions compared to observations for (top) maximum temperature in Helgoland, Germany and (bottom) precipitation in Madrid, Spain.

where f is parameterised by a neural network which is optimised by gradient descent, and Bayesian methods specifying a distribution over f. Conditional neural processes (CNPs; Garnelo et al., 2018) combine the advantages of these approaches, using neural networks to parameterise a stochastic process over the output variable, in this case either temperature or precipitation. Various improvements to the original CNP model have been suggested, for example the inclusion of a global latent variable

120 (Garnelo et al., 2018) and attention to improve the flexibility of the model and prevent under-fitting (Kim et al., 2019). In the context of spatial data such as downscaling predictions, a desirable characteristic of predictions are that they are translation-equivariant, that is the model makes identical predictions if data are spatially translated. The convCNP model (Gordon et al., 2019) applied here builds this equivariance into the CNP model. Throughout this study, we refer to the model developed in this section as 'the convCNP model'.

125 **2.1.1 Training**

The convCNP models are trained by minimising the average negative log-likelihood. For temperature, this is given by

$$NLL_{temp} = -\frac{1}{N}\sum_{i=1}^{N} log[\mathcal{N}(y_i|\mu_i(\mathbf{x}_i, Z, \mathbf{e}_i), \sigma_i(\mathbf{x}_i, Z, \mathbf{e}_i))$$

where y_i is the observed value, $\mathcal{N}(y_i; \mu_i, \sigma_i)$ denotes a Gaussian distribution over y with mean μ_i and variance σ_i^2. These parameters $\theta(\mathbf{x}_i, Z, \mathbf{e}_i) = \{\mu_i(\mathbf{x}_i, Z, \mathbf{e}_i), \sigma_i^2(\mathbf{x}_i, Z, \mathbf{e}_i)\}$ are generated by the model at each location $\mathbf{x}_i$ and use topography $\mathbf{e}_i$. N

130 is the total number of target locations. For precipitation, the negative log likelihood is given by

$$NLL_{precip} = -\frac{1}{N}\sum_{i=1}^{N}[r_i(ln(\rho_i(\mathbf{x}_i, Z, \mathbf{e}_i)) + ln(\Gamma(y_i|\alpha_i(\mathbf{x}_i, Z, \mathbf{e}_i), \beta_i(\mathbf{x}_i, Z, \mathbf{e}_i))) + (1-r_i)ln(1-\rho_i(\mathbf{x}_i, Z, \mathbf{e}_i))]$$

where r_i is a Bernoulli random variable describing whether precipitation was observed at the ith target location, y_i is the observed precipitation, ρ_i parameterises the predicted Bernoulli distribution and $\Gamma(y_i; \alpha_i, \beta_i)$ is a Gamma distribution with shape parameter α_i and scale parameter β_i. Here $\theta(\mathbf{x}_i, Z, \mathbf{e}_i) = \{\rho_i(\mathbf{x}_i, Z, \mathbf{e}_i), \alpha_i(\mathbf{x}_i, Z, \mathbf{e}_i), \beta_i(\mathbf{x}_i, Z, \mathbf{e}_i)\}$.

135 Weights are optimised using Adam (Kingma and Ba, 2014), with the learning rate set to 5×10^{-4}. Each model is trained for 100 epochs on 456 batches of 16 days each, using early stopping with a patience of 10 epochs.

2.2 Experiments and datasets

Having addressed the first aim in developing the convCNP model, we next evaluate model performance via three experiments. The first experiment compares the convCNP model to an ensemble of existing downscaling methods following a standardised
140 experimental protocol. In contrast to the convCNP model, these methods are unable to make predictions at locations where training data are not available. In the second experiment, we assess the performance of the convCNP model at these unseen locations compared to a baseline constructed by interpolating single-site models. Finally, ablation experiments are performed to quantify the impact of including sub-grid scale topographic information on performance.

2.2.1 Experiment 1 - baseline comparison

145 ConvCNP model performance is first compared to strong baseline methods taken from the VALUE experimental protocol. VALUE (Maraun et al., 2015) provides a standardised suite of experiments to evaluate new downscaling methods, together with data benchmarking the performance of existing methods. In the VALUE 1a experiment, each downscaling method predicts the maximum temperature and daily precipitation at 86 stations across Europe (Figure 2), given gridded data from the ERA-Interim reanalysis (Dee et al., 2011). These stations are chosen as they offer continuous, high fidelity data over the training and held
150 out test periods and represent multiple different climate regimes (Gutiérrez et al., 2019). Data is taken from 1979-2008, with five-fold cross validation used over six-year intervals to produce a 30 year time-series.

The convCNPs are trained to predict maximum temperature and precipitation at these 86 VALUE stations given the ERA Interim grids over Europe. Station data are taken from the European Climate Assessment Dataset (Klein Tank et al., 2002). These grids are restricted to points between 35 and 72 degrees latitude and -15 to 40 degrees longitude. The VALUE experiment
155 protocol does not specify which predictors are used in the downscaling model (i.e which gridded variables are included in Z). Based on the predictors used by methods in the baseline ensemble, winds, humidity and temperature are included at multiple levels together with time, latitude, longitude and invariant fields. Predictors are summarised in Table 1.

For the sub-grid scale information for input into the final MLP, the point measurement of three products is provided at each station. True station elevation is taken from the Global Multi-resolution Terrain Elevation Dataset (USGS, 2010). This is
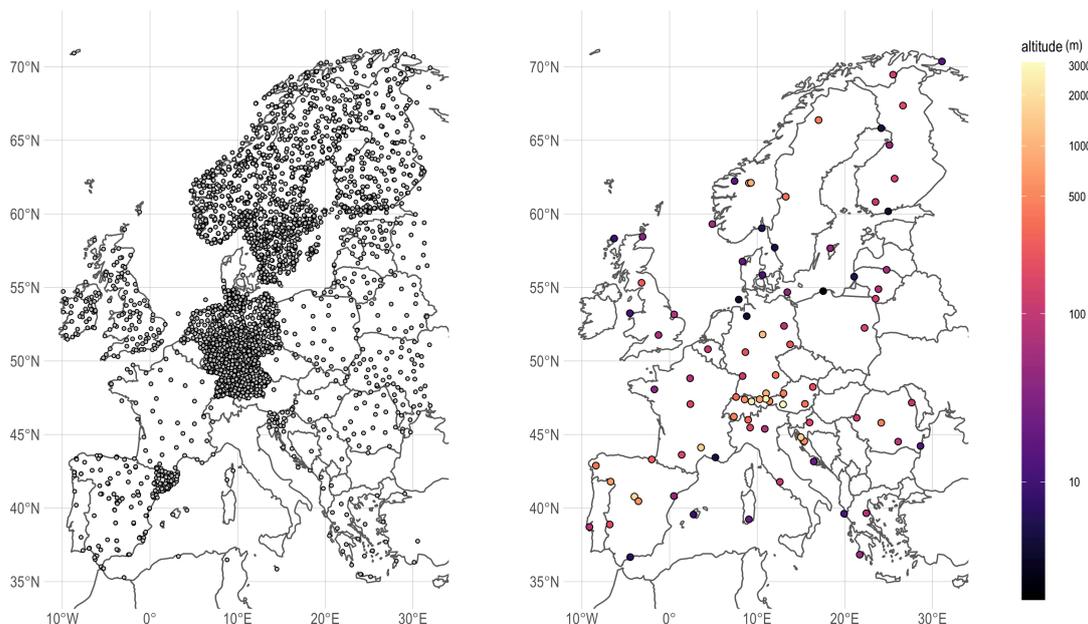
Figure 3. Locations of ECA&D (left) and VALUE (right) stations, with altitude shaded.

160   provided to the model together with the difference between the ERA-Interim gridscale resolution elevation and true elevation. Finally, topographic prominence is quantified using the ALOS Global mTPI (Theobald et al., 2015).

    Results of the convCNP model are compared to all available PP models in the VALUE ensemble, a total of 16 statistical models for precipitation and 23 for maximum temperature. These models comprise a range of techniques including analogs, multiple linear regression, generalised multiple linear regression and genetic programming. For a complete description of all

165   models included in the comparison, see Appendix A.

2.2.2   Experiment 2 - performance at unseen locations

We next quantify model performance at unseen locations compared to an interpolation baseline. The convCNP models are retrained using station data from the European Climate Assessment Dataset (ECA&D), comprising 3010 stations for precipitation and 3047 stations for maximum temperature (Figure 2). The 86 VALUE stations are held out as the validation set, testing

170   the model performance at both unseen times and locations.

    As existing downscaling models are unable to handle unseen locations, it is necessary to construct a new baseline. A natural baseline for this problem is to construct individual models for each station using the training set, use these to make predictions at future times, and then interpolate to get predictions at the held out locations. For the single-station models, predictors are taken from ERA-Interim data at the closest gridbox, similar to Gutiérrez et al. (2013). Multiple linear regression is used for maximum

175   temperature. For precipitation, occurrence is modelled using logistic regression, and accumulation using a generalised linear

Predictor	Level	Description
	Surface	
TMAX	surface	Maximum temperature
TMEAN	surface	Mean temperature
U10	surface	Northward wind
V10	surface	Eastward wind
Pr	surface	Accumulated precipitation
	Upper level	
Q	850hPa, 700hPa, 500hPa	Specific humidity
TA	850hPa, 700hPa, 500hPa	Temperature
UA	850hPa, 700hPa, 500hPa	Northward wind
VA	850hPa, 700hPa, 500hPa	Eastward wind
	Invariant	
ASO	surface	Angle of sub-gridscale orography
ANSO	surface	Anisotropy of sub-grid scale orography
FSO	surface	Standard deviation of filtered subgrid orography
SDO	surface	Standard deviation of orography
GSFC	surface	Geopotential
LAT	surface	Latitude
LON	surface	Longitude
	Temporal	
Time	-	Day of year, transformed as $(cos(time), sin(time))$

Table 1. Gridded predictors from ERA-Interim reanalysis included in Z.

model with gamma error distribution, similar to San-Martín et al. (2017). These methods are chosen as they are amongst the best-performing methods of the VALUE ensemble for each variable (Gutiérrez et al., 2019).

Following techniques used to convert station observations to gridded datasets (Haylock et al., 2008), predictions at these known stations in the future time period are made by first interpolating monthly means (totals) for temperature (precipitation) using a thin-plate spline, then using a GP to interpolate the anomalies (fraction of the total value). All interpolation is three dimensional over longitude, latitude and elevation. Throughout the results section, this model is referred to as the GP-baseline.

2.2.3 Experiment 3 - topography ablation

Finally, the impact of topography on predictions is quantified. Experiment 2 is repeated three times with different combinations of topographic data fed into the final MLP (step 3 in Figure 1): no topographic data, elevation and elevation difference only and mTPI only.

Means		
Metric	Variables	Description
mb	T_{max}, Precip	mean bias
sp	T_{max}, Precip	Spearman correlation between observed and predicted timeseries
MAE	T_{max}, Precip	Mean absolute error
R01	Precip	Relative wet day frequency (predicted precipitation days: observed precipitation days.
SDII	Precip	Mean wet day precipitation
Extremes		
Metric	Variables	Description
98P	T_{max}, precip	Bias in the 98th percentile.
R10	Precip	Relative frequency of days with precipitation greater than 10mm

Table 2. Evaluation metrics.

2.3 Evaluation metrics

A selection of standard climate metrics are chosen to assess model performance over the evaluation period, quantifying the representation of mean properties and extreme events (Table 2). Metrics are chosen based on those reported for the VALUE baseline ensemble (Gutiérrez et al., 2019; Widmann et al., 2019; Maraun et al., 2019; Hertig et al., 2019).

190   Comparison to these metrics requires generating a timeseries of values from the distributions predicted by the convCNP model. For temperature, this is generated by taking the mean of the predicted distribution for mean metrics, and sampling is used to complete the extreme metrics. For precipitation, a day is first classified as wet if $\rho \geq 0.5$ or dry if $\rho < 0.5$. For wet days, accumulations are generated by taking the mean of the gamma distribution for mean metrics or sampling for extreme metrics.

3 Results: baseline comparison (experiment 1)

195 The convCNP model outperforms all VALUE baselines on median mean absolute error (MAE) and spearman correlation for both maximum temperature and precipitation. Comparisons of convCNP model performance at the 86 VALUE stations to each model in the VALUE baseline ensemble are shown in Figure 4. The low MAE and high spearman correlation indicate that the model performs well at capturing day-to-day variability.

For maximum temperature, the mean bias is larger than baseline models at many stations, with interquartile range -0.02C
200 to 0.08C. This is a direct consequence of training a global model as opposed to individual models to each station which will trivially correct the mean (Maraun and Widmann, 2018). Though larger than baseline models, this error is still small for a majority of stations. Similarly for precipitation, though mean biases are larger than many of the VALUE models, the
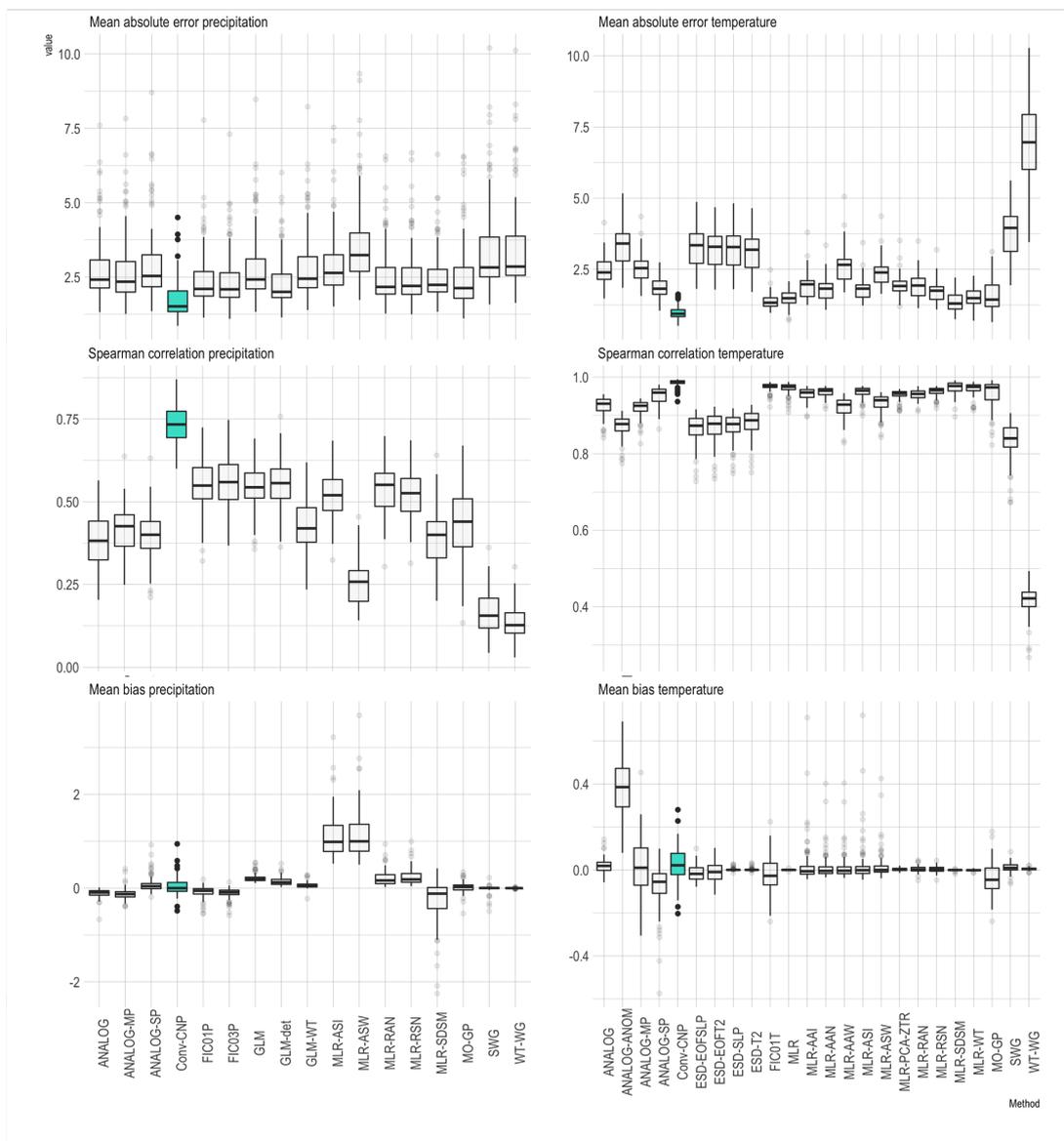
Figure 4. Comparison of the convCNP model to VALUE ensemble baselines for mean metrics, with the convCNP model shaded in blue.
Each box summarises performance for one model in the ensemble over the 86 training stations on the held out validation data.

interquartile range is just -0.07mm to 0.12mm. For precipitation, the bias in convCNP relative wet day frequency (R01) and mean wet day precipitation (SDII) are comparable to the best models in the VALUE ensemble (not shown).

205 When downscaling GCM output for impact studies, it is of particular importance to accurately reproduce extreme events (Katz and Brown, 1992). In line with previous work comparing the VALUE baselines (Hertig et al., 2019), an extreme event is defined to be a value greater than the 98th percentile of observations. Comparisons of biases in the 98th percentile of maximum temperature and precipitation are shown in Figure 5. The convCNP performs similarly to the best baselines, with a median bias of -0.02C for temperature and -2.04mm for precipitation across the VALUE stations. R10 biases are comparable to baselines, 210 with a median bias of just -0.003mm.

4   Results: performance at unseen locations (experiment 2)

The convCNP model outperforms the GP-baseline at unseen stations. Results for MAE, spearman correlation and mean bias are shown in Figure 6. For maximum temperature, the convCNP model gives small improvements over the baseline model, with spearman correlations of 0.99 (0.98) and MAE of 1.19C (1.35C) for the convCNP (GP baseline). Importantly, large 215 outliers (>10C) in the baseline MAE are not observed in the convCNP predictions. Figure 7 shows the spatial distribution of MAE for the convCNP and GP-baseline together with the difference in MAE between the two models. This demonstrates that stations with high MAE in the convCNP model are primarily concentrated in the complex topography of the European Alps. The GP-baseline model displays large MAE not only in the Alps, but also at other locations for example in Spain and France. The convCNP improves predictions at 82 out of the 86 stations.

220 Repeating this analysis for precipitation, the convCNP model gives substantial improvement over the baseline for MAE and spearman correlation. Spearman correlations are 0.57 (0.20) and MAE 2.10mm (2.71mm) for convCNP (GP-baseline). In contrast to maximum temperature, there is no clear link between topography and MAE, though again convCNP predictions have large MAE for multiple stations located in the Alps. The convCNP model improves on baseline predictions at 80 out of 86 stations.

225 Comparisons between models for extreme metrics are shown in Figure 8. For maximum temperature, the convCNP has slightly lower absolute 98th percentile bias than the baseline. For precipitation, errors are substantially lower, with median absolute 98th percentile bias of 4.90mm for convCNP compared to 22.92mm for the GP-baseline. The spatial distributions of 98th percentile bias for maximum temperature and precipitation predictions together with the difference in absolute bias are shown in Figure 9. For maximum temperature, the convCNP does not improve on the baseline at all stations. The GP-baseline 230 exhibits uniformly positive biases, while the convCNP model has both positive and negative biases. Improvements are seen through central and eastern Europe, while the convCNP performs comparatively poorly in Southern Europe and the British Isles. For precipitation, predictions have low biases across much of Europe for the convCNP, with the exception of in the complex terrain of the Alps. GP-baseline biases are negative throughout the domain. For this case, convCNP predictions have lower bias at 84 of the 86 validation stations.
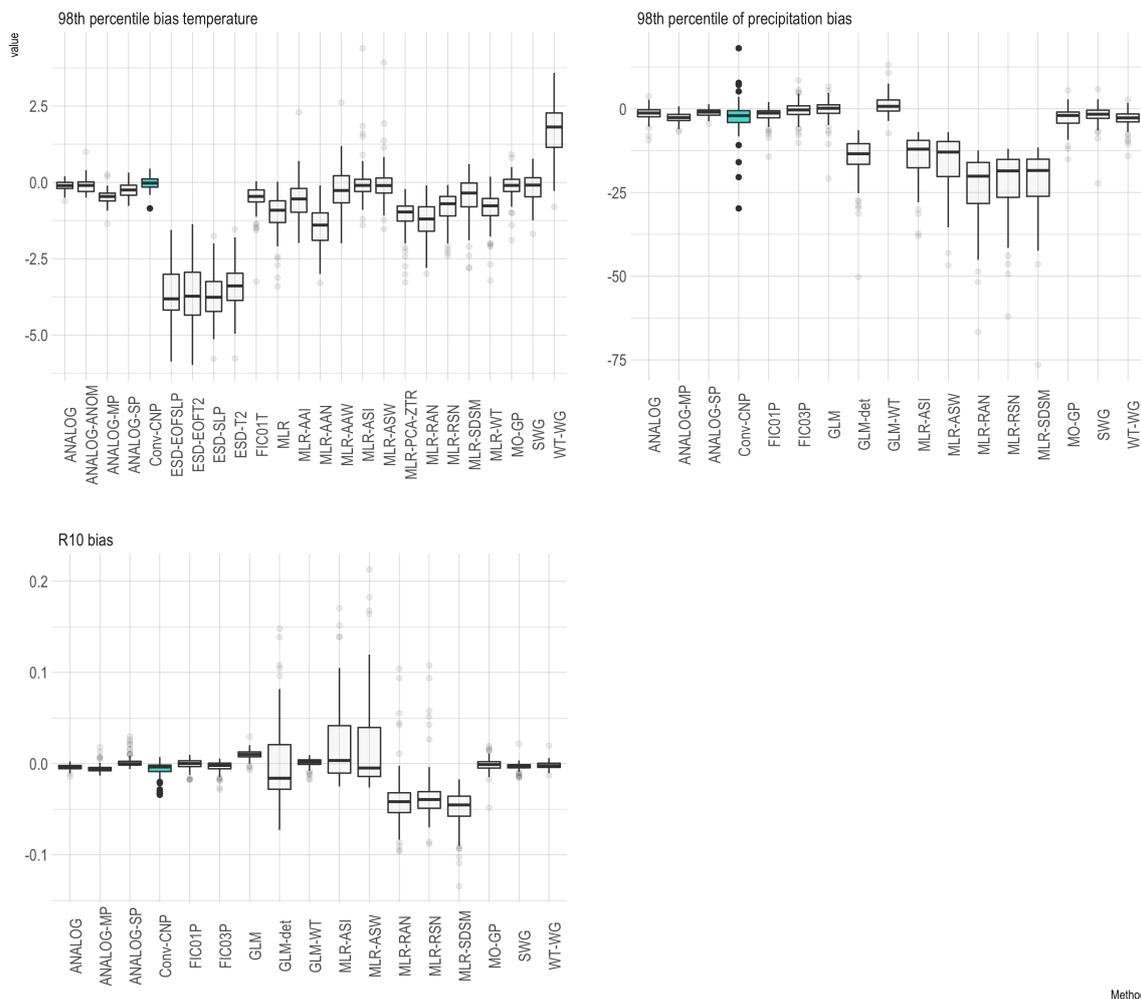
Figure 5. As for Figure 4, but for extreme metrics.
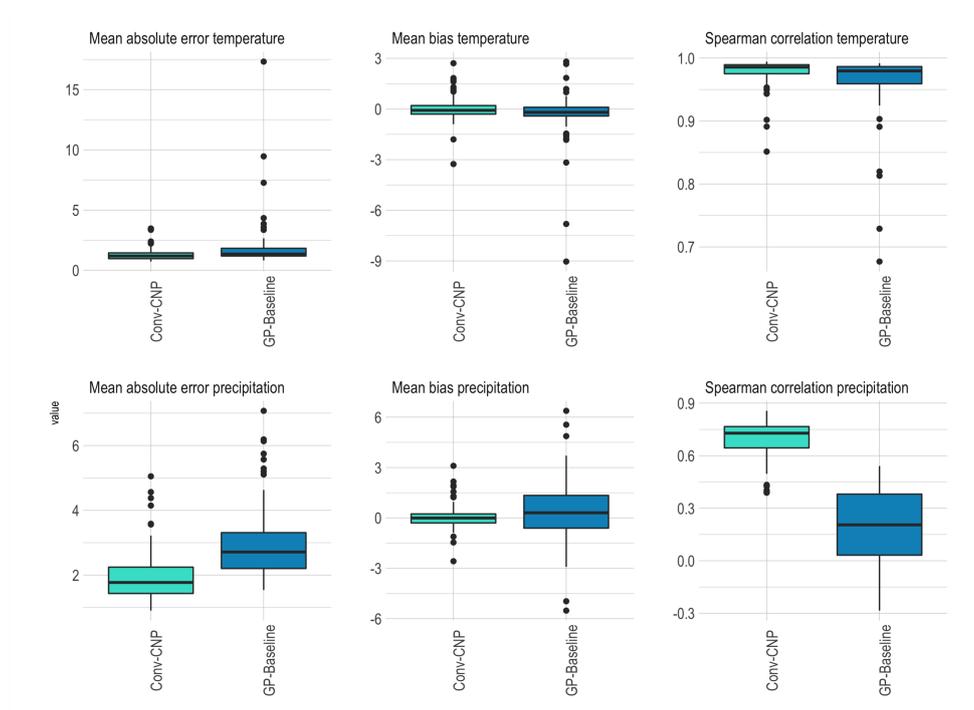
Figure 6. Comparison of the convCNP model to the GP-baseline. Boxes summarise model performance over the 86 held out VALUE stations for maximum temperature (top) and precipitation (bottom) for each of the mean metrics.

235   A limitation to the analysis of the standard climate metrics in Table 2 is that these only assess certain aspects of the predicted distribution. To assess the calibration of the models, we next examine the probability integral transform (PIT) values. The PIT value for a given prediction is defined as the CDF of the distribution predicted by the convCNP model evaluated at the true observed value. These values can be used to determine whether the model is calibrated by evaluating the PIT for every model prediction at the true observation, and plotting their distribution. If the model is properly calibrated, it is both necessary and
240   sufficient for this distribution to be uniform (Gneiting et al., 2007). PIT distributions for maximum temperature and wet-day precipitation are shown in Figure 10. For temperature, the model is well calibrated overall, although the predicted distributions are often too narrow, as demonstrated by the peaks around zero and one indicating that the observed value falls outside the predicted normal distribution. Calibration of the precipitation model is poorer overall. The peak in PIT mass around zero indicates that this model often over predicts rainfall accumulation. Performance varies between individual stations for both
245   temperature and precipitation, with examples of PIT distributions for both well- and poorly-calibrated stations shown in Figure 10.
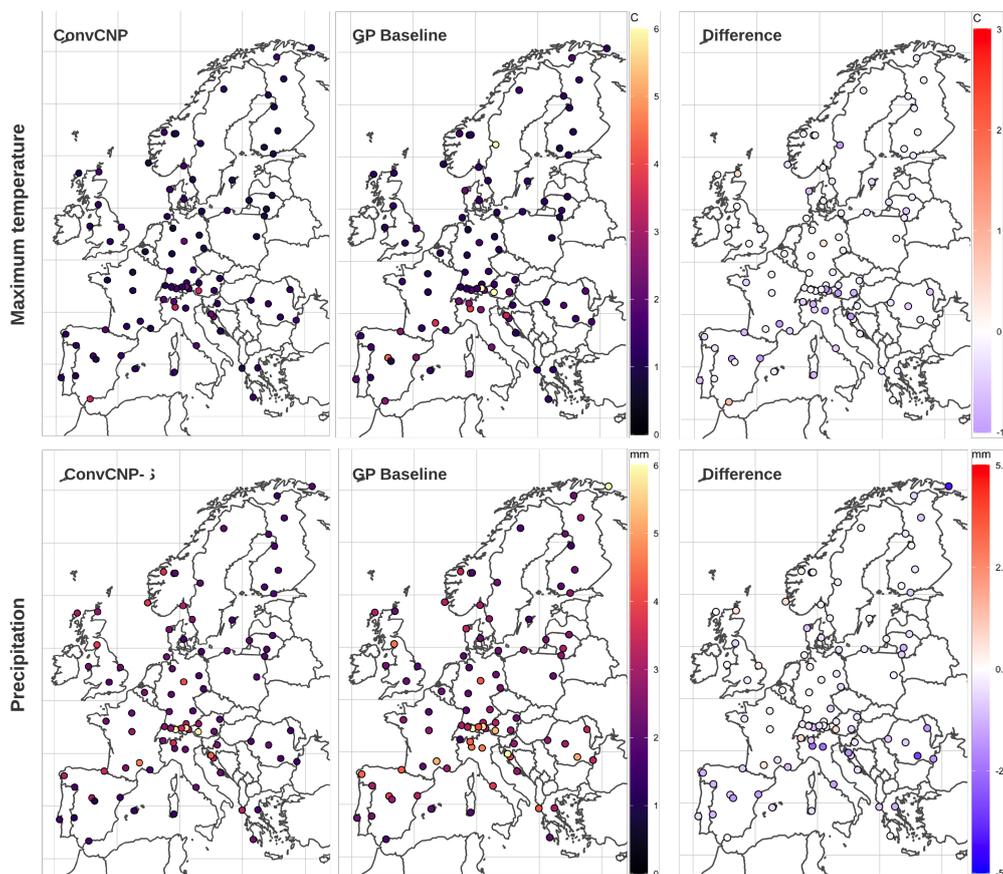
Figure 7. Spatial distribution of mean absolute error for convCNP (right), GP baseline (centre) and convCNP - GP baseline (left). Maximum temperature (precipitation) is shown on the top (bottom) row. All panels show results for each of the 86 held out VALUE stations.

5   Results: topography ablation (experiment 3)

Results of the topography ablation experiment are shown in Figure 11 (mean metrics) and Figure 12 (extreme metrics). These figures compare the performance on each metric between the convCNP model with all topographic predictors, and convCNP models trained with no topography, elevation and elevation difference only and mTPI only.

For maximum temperature, inclusion of topographic information improves MAE, mean bias and spearman correlation. Models including only mTPI or no topographic predictors have a number of stations with very large MAE, exceeding 10C at several stations. Unsurprisingly, these stations are found to be located in areas of complex topography in the Alps (not shown). Including elevation both decreases the median MAE and corrects errors at these outliers, with further improvement
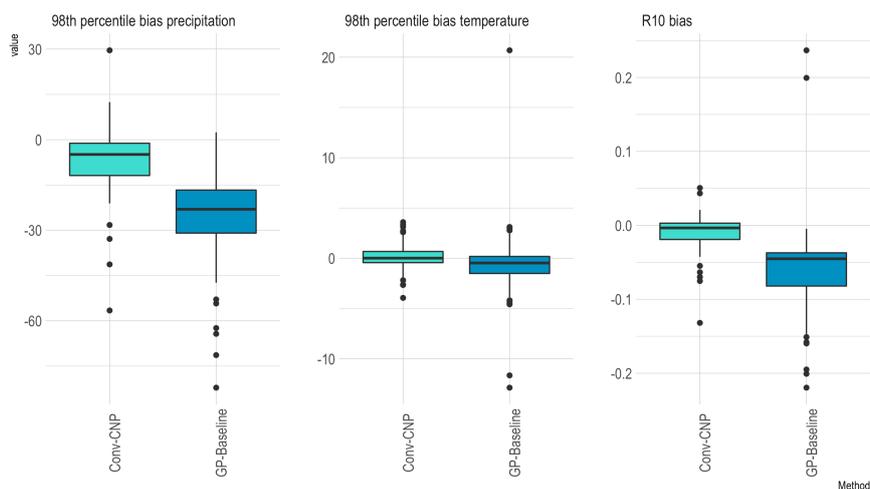
Figure 8. As for Figure 6, but for P98 and R10 biases.

observed with mTPI added. A similar pattern is seen for mean bias. More modest improvements are seen for precipitation, though inclusion of topographic data does result in slightly improved performance.

For maximum temperature, inclusion of topographic data results in reduced 98th percentile bias. This is primarily as a result of including elevation and elevation difference data, with limited benefit derived from the inclusion of mTPI. In contrast, for precipitation, models with topographic correction perform worse than the elevation agnostic model for both 98th percentile and R10 biases. This reduced performance for precipitation may result from overfitting.

6 Discussion and conclusion

This study demonstrated the successful application on convCNPs to statistical downscaling of temperature and precipitation. The convCNP model performs well compared to strong baselines from the VALUE ensemble on both mean and extreme metrics. For both variables the convCNP model outperforms an interpolation based baseline. Inclusion of sub-grid-scale topographic information is shown to improve model performance for mean and extreme metrics for maximum temperature, and mean metrics for precipitation.

The convCNP model has several advantages over existing methods beyond performance on these metrics. The most important of these is that the trained model can be used to predict values at any target point, regardless of whether this point is included in the training set. In Sect. 2, we demonstrated that the convCNP model outperforms interpolation of single-site downscaling models. This has application in developing local projections for climate impact studies. A second useful feature is that the convCNP takes raw grid data as input, without the need for feature engineering. This contrasts with many of the
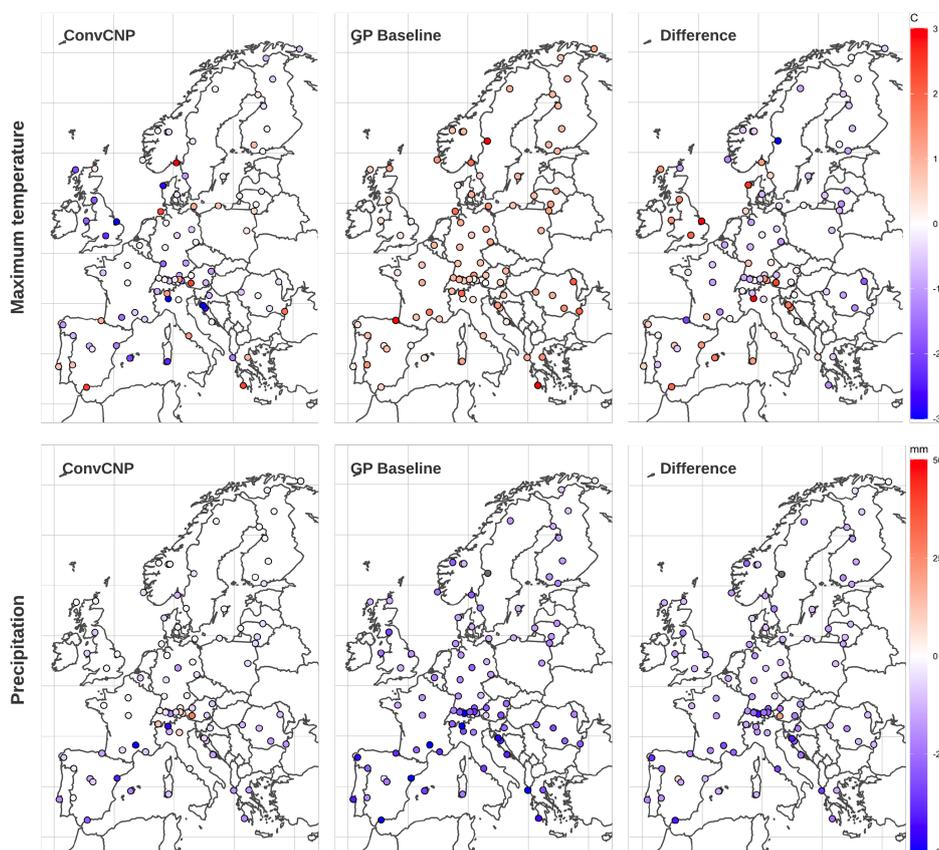
Figure 9. As for Figure 7, but for P98 biases. Here, the difference panels quantify the difference in absolute bias $|P98_{convCNP}| - |P98_{GP-Baseline}|$. Negative (positive) values indicate that the convCNP (GP-Baseline) has better performance.

baseline methods, which rely on the identification of relevant circulation analogs. Although only temperature and precipitation are considered in this study, the model is easily applied to any climate variable with available station observations, for example windspeed.

275   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
280   (Baño-Medina et al., 2020) and precipitation (Vandal et al., 2019), and motivates continued efforts to develop and benchmark such models for application in climate impact studies.
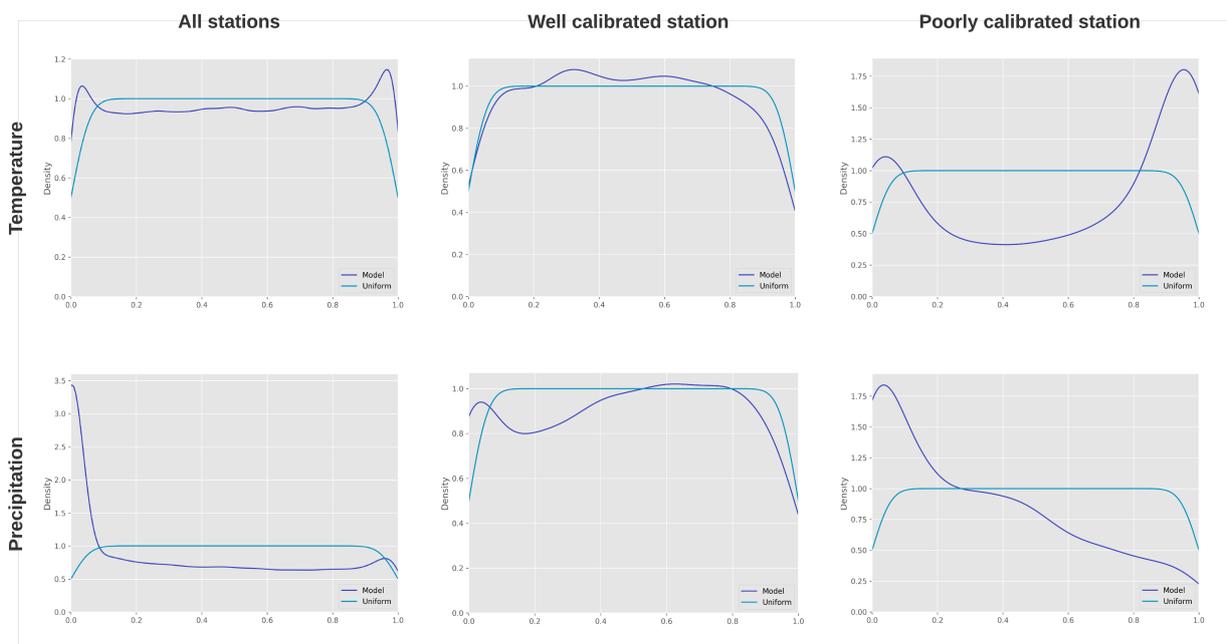
Figure 10. KDEs showing probability integral transforms for model predictions compared to a uniform distribution to assess calibration for temperature (top) and precipitation (bottom). For temperature PIT plots are shown for all values (left), a station where the model is well calibrated (Braganca, Portugal; centre) and a station where the model is poorly calibrated (Gospic, Croatia; right). Similarly for precipitation, PIT plots are shown for all values (left), a station where the model is well calibrated (Stornoway, UK; centre) and a station where the model is poorly calibrated (Sondankyla, Finland; right).

Several areas remain for future work. Representation of certain metrics, notably precipitation extremes requires further improvement, particularly in areas with complex topography. The topography ablation experiments demonstrate that the convCNP P98 bias increases with the inclusion of topographic information. Dynamically, this is likely due to local flow effects such as Föhn winds (Gaffin, 2007; Basist et al., 1994), which depend on the incident angle of the background flow. A possible explanation for this is that the MLP is insufficient to model these effects. Further experimentation with adding a second CNN to capture the sub-grid-scale processes, and possibly conditioning predictions of this model on local flow is left as a topic for future research.

Another avenue for improving model performance would be to change the distribution predicted by the convCNP. Model calibration results presented in Section 4 indicate that the temperature downscaling model could be improved using a distribution with heavier tails. Precipitation model calibration requires improvement, with the model frequently under-predicting wet day accumulations. A possible explanation for this is that the left hand tail of the gamma distribution decays rapidly. For
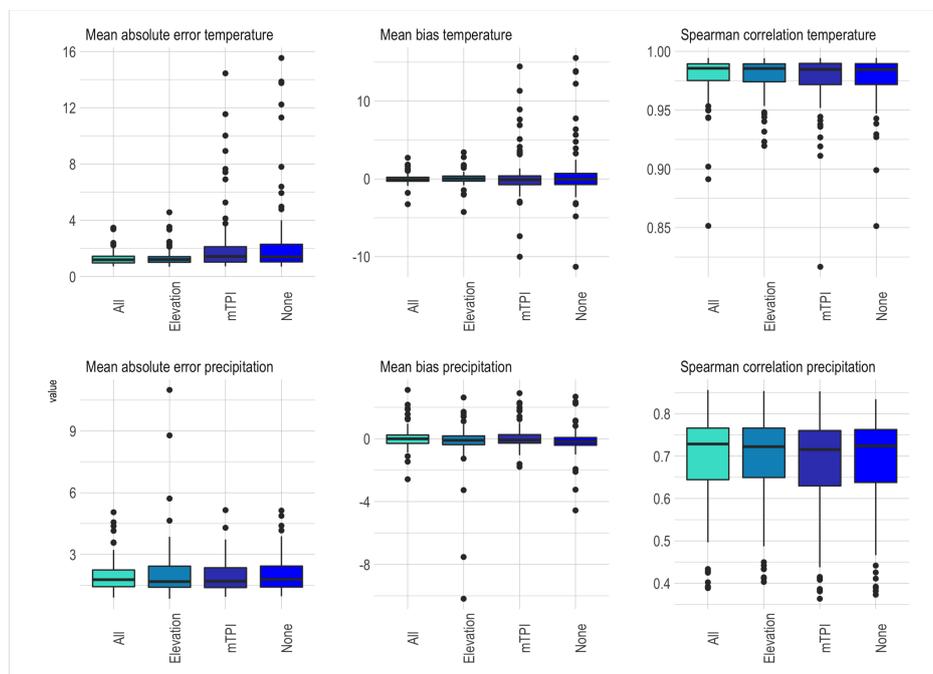
Figure 11. Comparison of model performance in the topography ablation experiment. The complete model (All) is compared to models with no topographic data (None), elevation and elevation difference only (Elevation) and mTPI only (mTPI). Boxes summarise model performance over the 86 held out VALUE stations for maximum temperature (top) and precipitation (bottom) for each of the mean metrics.

cases where the mode of the predicted distribution is greater than zero, small observed accumulations are therefore heavily penalised. Previous work has acknowledged that the Bernoulli-Gamma distribution used in this study is not realistic for all
295 sites (Vlček and Huth, 2009), and suggested that representation of precipitation extremes can be improved using a Bernoulli-Gamma-Gerneralised Pareto distribution (Ben Alaya et al., 2015; Volosciuk et al., 2017). Future work will explore improving the calibration of the downscaling models using mixture distributions and normalising flows (Rezende and Mohamed, 2015) to improve the calibration of the model. A further possibility for extending the convCNP model would be to explicitly incorporate time by building recurrence into the model (Qin et al., 2019; Singh et al., 2019).
300   The final aspect to consider is extending these promising results downscaling reanalysis data to apply to future climate simulations from GCMs. An in depth analysis of the convCNP model performance on seasonal and annual metrics would be beneficial in informing application to impact scenarios. A limitation in all PP downscaling techniques is that applying a transfer function trained on reanalysis data to a GCM makes the assumption that the predictors included in the context set are realistically simulated in the GCM (Maraun and Widmann, 2018). Future work will aim to address this issue through training a
305 convCNP model directly on RCM or GCM hindcasts available through projects such as EURO-CORDEX (Jacob et al., 2014).
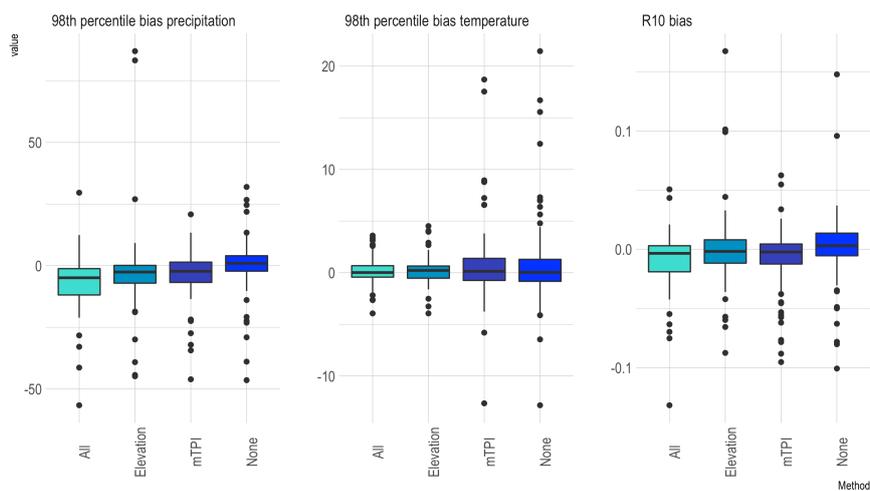
Figure 12. As for Figure 11, but for extreme metrics.

Code availability. Model code is available at https://github.com/anna-184702/convCNPClimate (DOI:10.5281/zenodo.4554603).

Author contributions. AV implemented the code, conducted the experiments and wrote the first draft. All authors designed the study and contributed to the analysis of results and final version of the paper.

Competing interests. The authors declare that they have no conflict of interest.

310   *Acknowledgements.* Anna Vaughan acknowledges the UKRI Centre for Doctoral Training in the Application of Artificial Intelligence to the study of Environmental Risks (AI4ER), led by the University of Cambridge and British Antarctic Survey, and studentship funding from Google DeepMind. Will Tebbutt is supported by Google DeepMind.

References

Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P.: Climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the Intergovernmental Panel on Climate Change, 2016.

Ayar, P. V., Vrac, M., Bastin, S., Carreau, J., Déqué, M., and Gallardo, C.: Intercomparison of statistical and dynamical downscaling models under the EURO-and MED-CORDEX initiative framework: present climate evaluations, Climate Dynamics, 46, 1301–1329, 2016.

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.

Basist, A., Bell, G. D., and Meentemeyer, V.: Statistical relationships between topography and precipitation patterns, Journal of climate, 7, 1305–1315, 1994.

Ben Alaya, M. A., Chebana, F., and Ouarda, T. B.: Probabilistic multisite statistical downscaling for daily precipitation using a Bernoulli–generalized pareto multivariate autoregressive model, Journal of Climate, 28, 2349–2364, 2015.

Benestad, R. E., Chen, D., Mezghani, A., Fan, L., and Parding, K.: On using principal components to represent stations in empirical–statistical downscaling, Tellus A: Dynamic Meteorology and Oceanography, 67, 28 326, 2015.

Bevacqua, E., Maraun, D., Hobæk Haff, I., Widmann, M., and Vrac, M.: Multivariate statistical modelling of compound events via pair-copula constructions: analysis of floods in Ravenna (Italy), Hydrology and Earth System Sciences, 21, 2701–2723, 2017.

Bhardwaj, A., Misra, V., Mishra, A., Wootten, A., Boyles, R., Bowden, J., and Terando, A. J.: Downscaling future climate change projections over Puerto Rico using a non-hydrostatic atmospheric model, Climatic Change, 147, 133–147, 2018.

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

Cannon, A. J., Piani, C., and Sippel, S.: Bias correction of climate model output for impact models, Elsevier, 2020.

Casanueva, A., Herrera, S., Fernández, J., and Gutiérrez, J. M.: Towards a fair comparison of statistical and dynamical downscaling in the framework of the EURO-CORDEX initiative, Climatic Change, 137, 411–426, 2016.

Chollet, F.: Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251–1258, 2017.

Dee, D. P., Uppala, S. M., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d. P., et al.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Quarterly Journal of the royal meteorological society, 137, 553–597, 2011.

Gaffin, D. M.: Foehn winds that produced large temperature differences near the southern Appalachian Mountains, Weather and forecasting, 22, 145–159, 2007.

Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S., and Teh, Y. W.: Neural processes, arXiv preprint arXiv:1807.01622, 2018.

Ghosh, S. and Mujumdar, P. P.: Statistical downscaling of GCM simulations to streamflow using relevance vector machine, Advances in water resources, 31, 132–146, 2008.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69, 243–268, 2007.

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.

350 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.

Gutiérrez, J. M., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., Rössler, O., Wibig, J., Wilcke, R., Kotlarski, S., et al.: An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment, International journal of climatology, 39, 3750–3785, 2019.

355 Hatfield, J. L. and Prueger, J. H.: Temperature extremes: Effect on plant growth and development, Weather and climate extremes, 10, 4–10, 2015.

Haylock, M., Hofstra, N., Klein Tank, A., Klok, E., Jones, P., and New, M.: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, Journal of Geophysical Research: Atmospheres, 113, 2008.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer
360   vision and pattern recognition, pp. 770–778, 2016.

Hertig, E. and Jacobeit, J.: A novel approach to statistical downscaling considering nonstationarities: application to daily precipitation in the Mediterranean area, Journal of Geophysical Research: Atmospheres, 118, 520–533, 2013.

Hertig, E., Maraun, D., Bartholy, J., Pongracz, R., Vrac, M., Mares, I., Gutiérrez, J. M., Wibig, J., Casanueva, A., and Soares, P. M.: Comparison of statistical downscaling methods with respect to extreme events over Europe: Validation results from the perfect predictor
365   experiment of the COST Action VALUE, International Journal of Climatology, 39, 3846–3867, 2019.

Höhlein, K., Kern, M., Hewson, T., and Westermann, R.: A Comparative Study of Convolutional Neural Network Models for Wind Field Downscaling, arXiv preprint arXiv:2008.12257, 2020.

Huth, R., Mikšovský, J., Štěpánek, P., Belda, M., Farda, A., Chládová, Z., and Pišoft, P.: Comparative validation of statistical and dynamical downscaling models on a dense grid in central Europe: temperature, Theoretical and Applied Climatology, 120, 533–553, 2015.

370 Jacob, D., Petersen, J., Eggert, B., Alias, A., Christensen, O. B., Bouwer, L. M., Braun, A., Colette, A., Déqué, M., Georgievski, G., et al.: EURO-CORDEX: new high-resolution climate change projections for European impact research, Regional environmental change, 14, 563–578, 2014.

Jacobeit, J., Hertig, E., Seubert, S., and Lutz, K.: Statistical downscaling for climate change projections in the Mediterranean region: methods and results, Regional environmental change, 14, 1891–1906, 2014.

375 Katz, R. W. and Brown, B. G.: Extreme events in a changing climate: variability is more important than averages, Climatic change, 21, 289–302, 1992.

Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, A., Rosenbaum, D., Vinyals, O., and Teh, Y. W.: Attentive neural processes, arXiv preprint arXiv:1901.05761, 2019.

King, A. D., Alexander, L. V., and Donat, M. G.: The efficacy of using gridded data to examine extreme rainfall characteristics: a case study
380   for Australia, International Journal of Climatology, 33, 2376–2387, 2013.

Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

Klein Tank, A., Wijngaard, J., Können, G., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., et al.: Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment, International Journal of Climatology: A Journal of the Royal Meteorological Society, 22, 1441–1453, 2002.

385 Maraun, D.: Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue, Journal of Climate, 26, 2137–2143, 2013.

Maraun, D. and Widmann, M.: Statistical downscaling and bias correction for climate research, Cambridge University Press, 2018.

Maraun, D., Wetterhall, F., Ireson, A., Chandler, R., Kendon, E., Widmann, M., Brienen, S., Rust, H., Sauter, T., Themeßl, M., et al.: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user, Reviews of geophysics, 48, 2010.

390 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.

Maraun, D., Shepherd, T. G., Widmann, M., Zappa, G., Walton, D., Gutiérrez, J. M., Hagemann, S., Richter, I., Soares, P. M., Hall, A., et al.: Towards process-informed bias correction of climate change simulations, Nature Climate Change, 7, 764–773, 2017.

Maraun, D., Huth, R., Gutiérrez, J. M., Martín, D. S., Dubrovsky, M., Fischer, A., Hertig, E., Soares, P. M., Bartholy, J., Pongrácz, R., et al.:

395 The VALUE perfect predictor experiment: evaluation of temporal variability, International Journal of Climatology, 39, 3786–3818, 2019.

Mehrotra, R. and Sharma, A.: A nonparametric nonhomogeneous hidden Markov model for downscaling of multisite daily rainfall occurrences, Journal of Geophysical Research: Atmospheres, 110, 2005.

Piani, C., Haerter, J., and Coppola, E.: Statistical bias correction for daily precipitation in regional climate models over Europe, Theoretical and Applied Climatology, 99, 187–192, 2010.

400 Qin, S., Zhu, J., Qin, J., Wang, W., and Zhao, D.: Recurrent attentive neural process for sequential data, arXiv preprint arXiv:1910.09323, 2019.

Raynaud, D., Hingray, B., Zin, I., Anquetin, S., Debionne, S., and Vautard, R.: Atmospheric analogues for physically consistent scenarios of surface weather in Europe and Maghreb, International Journal of Climatology, 37, 2160–2176, 2017.

Rezende, D. J. and Mohamed, S.: Variational inference with normalizing flows, arXiv preprint arXiv:1505.05770, 2015.

405 Ribalaygua, J., Torres, L., Pórtoles, J., Monjo, R., Gaitán, E., and Pino, M.: Description and validation of a two-step analogue/regression downscaling method, Theoretical and Applied Climatology, 114, 253–269, 2013.

Sachindra, D., Ahmed, K., Rashid, M. M., Shahid, S., and Perera, B.: Statistical downscaling of precipitation using machine learning techniques, Atmospheric research, 212, 240–258, 2018.

San-Martín, D., Manzanas, R., Brands, S., Herrera, S., and Gutiérrez, J. M.: Reassessing model uncertainty for regional projections of

410 precipitation with an ensemble of statistical downscaling methods, Journal of Climate, 30, 203–223, 2017.

Singh, G., Yoon, J., Son, Y., and Ahn, S.: Sequential Neural Processes, in: Advances in Neural Information Processing Systems, pp. 10 254–10 264, 2019.

Teutschbein, C. and Seibert, J.: Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods, Journal of hydrology, 456, 12–29, 2012.

415 Theobald, D. M., Harrison-Atlas, D., Monahan, W. B., and Albano, C. M.: Ecologically-relevant maps of landforms and physiographic diversity for climate adaptation planning, PLoS One, 10, e0143 619, 2015.

USGS: Global Multi-resolution Terrain Elevation Data, 2010.

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.

420 Vlček, O. and Huth, R.: Is daily precipitation Gamma-distributed?: Adverse effects of an incorrect use of the Kolmogorov–Smirnov test, Atmospheric Research, 93, 759–766, 2009.

Volosciuk, C. D., Maraun, D., Vrac, M., and Widmann, M.: A combined statistical bias correction and stochastic downscaling method for precipitation, Hydrology and Earth System Sciences, 21, 1693–1719, 2017.

Widmann, M., Bedia, J., Gutiérrez, J. M., Bosshard, T., Hertig, E., Maraun, D., Casado, M. J., Ramos, P., Cardoso, R. M., Soares, P. M.,
425   et al.: Validation of spatial variability in downscaling results from the VALUE perfect predictor experiment, International Journal of
        Climatology, 39, 3819–3845, 2019.

Wilby, R. L., Dawson, C. W., and Barrow, E. M.: SDSM?a decision support tool for the assessment of regional climate change impacts,
        Environmental Modelling & Software, 17, 145–157, 2002.

Zerenner, T., Venema, V., Friederichs, P., and Simmer, C.: Downscaling near-surface atmospheric fields with multi-objective Genetic Pro-
430   gramming, Environmental Modelling & Software, 84, 85–98, 2016.

Appendix A: VALUE ensemble methods

Table 3 summarises the baseline methods in the VALUE ensemble. This information is adapted from (Gutiérrez et al., 2019).

Model	Variables	Description
ANALOG	T_{max}, precip	Standard analog, no seasonal component (Gutiérrez et al., 2013)
ANALOG-ANOM	T_{max}	Analog with seasonal component (Ayar et al., 2016)
ANALOG-MP	T_{max}, precip	Analog method with seasonal component (Raynaud et al., 2017)
ANALOG-SP	T_{max}, precip	Analog method with seasonal component (Raynaud et al., 2017)
ESD-EOFSLP	T_{max}	Multiple linear regression (Benestad et al., 2015)
ESD-SLP	T_{max}	Multiple linear regression (Benestad et al., 2015)
ESD-T2	T_{max}	Multiple linear regression (Benestad et al., 2015)
FIC01P	T_{max}, precip	Two step analog method (Ribalaygua et al., 2013)
FIC03P	precip	Two step analog method (Ribalaygua et al., 2013)
GLM	precip	Generalised linear Model with log-canonical link function (San-Martín et al., 2017). Bernoulli error distribution for occurrence and gamma error distribution for accumulation. Predictions sampled from output distribution.
GLM-det	precip	As for GLM, predictions given as mean of output distribution (San-Martín et al., 2017).
GLM-WT	precip	As for GLM, conditioned on 12 weather types identified by k-means clustering (San-Martín et al., 2017).
MLR	T_{max}	Multiple linear regression using PCA for predictors (Gutiérrez et al., 2013).
MLR-AAI	T_{max}, precip	Multiple linear regression, annual training, anomaly data, inflation variance correction (Huth et al., 2015).
MLR-AAN	T_{max}, precip	Multiple linear regression, annual training, anomaly data, white-noise variance correction (Huth et al., 2015).
MLR-AAW	T_{max}, precip	Multiple linear regression, annual training, anomaly data, white-noise variance correction (Huth et al., 2015).
MLR-ASI	T_{max}, precip	Multiple linear regression, seasonal training, anomaly data, inflation variance correction (Huth et al., 2015).
MLR-ASW	T_{max}, precip	Multiple linear regression, seasonal training, anomaly data, white-noise variance correction (Huth et al., 2015).
MLR-PCA-ZTR	T_{max}	Multiple linear regression with s-mode principal component predictors (Jacobeit et al., 2014).
MLR-RAN	T_{max}, precip	Multiple linear regression, seasonal training, raw data, no variance correction (Huth et al., 2015).
MLR-RSN	T_{max}, precip	Multiple linear regression, seasonal training, raw data, no variance correction (Huth et al., 2015).
MLR-SDSM	T_{max}, precip	Single-site multiple linear regression using the statistical downscaling method (Wilby et al., 2002).
MLR-WT	T_{max}	As for MLR, but conditioned on weather types defined using k-means clustering (Gutiérrez et al., 2013).
MO-GP	T_{max}, precip	Multi-objective genetic programming (Zerenner et al., 2016).
SWG	T_{max}, precip	Two-step vectorised generalised linear models (Ayar et al., 2016).
WT-WG	T_{max}, precip	Distributional fitting based on weather types selected using k-means clustering (Gutiérrez et al., 2013).

Table A1. Summary of models included in the VALUE ensemble.