

RESEARCH ARTICLE

Multi-objective downscaling of precipitation time series by genetic programming

Tanja Zerenner^{1,2}  | Victor Venema¹  | Petra Friederichs¹  |
Clemens Simmer¹ 

¹Meteorological Department of the
Institute for Geosciences, University of
Bonn, Bonn, Germany

²Department of Mathematics, College of
Engineering, Mathematics and Physical
Sciences, University of Exeter, Exeter,
United Kingdom

Correspondence

Tanja Zerenner, Department of
Mathematics, University of Sussex,
Falmer, Brighton BN1 9QH, United
Kingdom.

Email: t.zerenner@sussex.ac.uk

Funding information

CRC/TR32: Patterns in Soil-Vegetation-
Atmosphere Systems: Monitoring,
Modelling and Data Assimilation; funded
by the German Research Foundation
(Deutsche Forschungsgemeinschaft DFG)

Abstract

We use symbolic regression to estimate daily precipitation amounts at six stations in the Alpine region from a global reanalysis. Symbolic regression only prescribes the set of mathematical expressions allowed in the regression model, but not its structure. The regression models are generated by genetic programming (GP) in analogy to biological evolution. The two conflicting objectives of a low root-mean-square error (RMSE) and consistency in the distribution between model and observations are treated as a multi-objective optimization problem. This allows us to derive a set of downscaling models that represents different achievable trade-offs between the two conflicting objectives, a so-called Pareto set. Our GP setup limits the size of the regression models and uses an analytical quotient instead of a standard or protected division operator. With this setup we obtain models that have a generalization performance comparable with generalized linear regression models (GLMs), which are used as a benchmark. We generate deterministic and stochastic downscaling models with GP. The deterministic downscaling models with low RMSE outperform the respective stochastic models. The stochastic models with low IQD, however, perform slightly better than the respective deterministic models for the majority of cases. No approach is uniquely superior. The stochastic models with optimal IQD provide useful distribution estimates that capture the stochastic uncertainty similar to or slightly better than the GLM-based downscaling.

KEYWORDS

genetic programming, machine learning, Pareto optimality, stochastic downscaling

1 | INTRODUCTION

1.1 | Empirical-statistical downscaling

General circulation models (GCM) simulate the climate system under past, present and future conditions. GCMs

provide valuable information about the future climate, but at spatial resolutions that are often too coarse for impact studies for two reasons: First, GCM simulations represent values for large grid boxes of up to 100 km and more, and are thus not directly comparable to local station observations. Second, effects of subgrid-scale

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors *International Journal of Climatology* published by John Wiley & Sons Ltd on behalf of Royal Meteorological Society.

heterogeneity, such as local and regional topography or land-sea distribution, are not resolved in GCMs.

One approach to bridge the scale gap between the coarse climate model output and the requirements of impact modellers is empirical-statistical downscaling (Benestad *et al.*, 2008; Maraun and Widmann, 2018), which relates a local variable, such as locally observed precipitation at a station, to the larger-scale atmospheric state provided by a GCM or regional climate model (RCM). Near-surface temperature and precipitation are the most considered climate variables in empirical-statistical downscaling due to their importance in impact studies. Its high spatial and temporal variability and non-Gaussian distribution over a wide range of scales make precipitation downscaling particularly challenging (e.g., Maraun *et al.*, 2010).

Common approaches and methods in empirical-statistical downscaling include regression methods/transfer functions, weather typing methods and stochastic weather generators (Hewitson and Crane, 1996; Wilby and Wigley, 1997). Regression establishes a statistical relationship, which estimates the conditional expectation of the local predictand given the larger-scale predictors. Regression methods in downscaling comprise linear techniques, such as multiple linear regression (e.g., Huth, 2002; Gutiérrez *et al.*, 2013; Huth *et al.*, 2015) or generalized linear models (GLM) (e.g., Chandler and Wheeler, 2002; Abaurrea and Asín, 2005; San-Martín *et al.*, 2017), and non-linear techniques such as artificial neural networks (e.g., Schoof and Pryor, 2001; Coulibaly *et al.*, 2005; Huth *et al.*, 2015; Baño-Medina *et al.*, 2020) or genetic programming (e.g., Coulibaly, 2004; Hashmi *et al.*, 2011; Sachindra and Kanae, 2019). Weather typing approaches relate the local variable to the occurrence of a particular weather class defined on the larger scale (e.g., Zorita *et al.*, 1995; Vrac *et al.*, 2007a; Cheng *et al.*, 2011). Weather classes can be derived objectively for instance through principal component analysis or subjectively using for instance established circulation classification schemes for the region of interest. Weather generators (WGs) are stochastic models which aim to reproduce the statistics such as mean, variance and auto-correlation of the local observations (Richardson, 1981; Wilks and Wilby, 1999). A common first step in a WG is to model precipitation occurrence by a Markov chain. Precipitation intensity is then obtained in a second step by sampling from a suitable (e.g., gamma or exponential) distribution. The remaining variables of interest, most commonly temperature, are modelled by an auto-regressive model conditioned on precipitation. WGs are not per se downscaling methods, but are frequently used as such by conditioning the WG parameters on the larger-scale atmospheric state (e.g., Wilby *et al.*, 2002; Kilsby *et al.*, 2007; Keller *et al.*, 2017).

In the context of empirical-statistical downscaling, genetic programming (GP) is typically used to perform a

symbolic regression (e.g., Coulibaly, 2004; Hashmi *et al.*, 2011; Sachindra and Kanae, 2019). In symbolic regression mathematical expressions, variables and constants are flexibly combined to build regression models. Optimization proceeds analogous to biological evolution (Koza, 1992; Banzhaf *et al.*, 1997; Poli *et al.*, 2008), that is, models are evolved over several generations based on the principle of the survival of the fittest. Starting from an initial population of randomly generated models, each subsequent generation is generated by modifying models of the previous generation. The better a model performs with respect to a predefined fitness measure (for symbolic regression typically the RMSE) the more likely it will contribute to the new generation.

Several studies have intercompared empirical-statistical downscaling techniques (e.g., Frost *et al.*, 2011; Gutmann *et al.*, 2014; San-Martín *et al.*, 2017; Gutiérrez *et al.*, 2019). These studies usually do not identify one best technique, but rather provide users with guidance on which techniques to choose under what conditions. The most comprehensive intercomparison of empirical-statistical downscaling methods to date was initiated by the European Cooperation in Science & Technology (COST) Action VALUE on *Validating and Integrating Downscaling Methods for Climate Change Research* (Maraun *et al.*, 2015). A summary of the results of the over 50 contributing methods can be found in Gutiérrez *et al.* (2019), an in depth evaluation with respect to extremes, spatial variability, temporal variability and atmospheric processes in Hertig *et al.* (2019); Widmann *et al.* (2019); Maraun *et al.* (2019) and Soares *et al.* (2019). The COST-VALUE intercomparison covers variants of the most common empirical-statistical downscaling approaches, including regression, weather typing and analog methods, weather generators as well as bias correction and quantile mapping techniques. The only GP-based approach contributed is an earlier variant of the multi-objective GP method used in the present study. Nevertheless, several studies have applied evolutionary methods to downscaling, the majority of which minimize the RMSE or a similar measure between downscaling estimate and reference/observation (Coulibaly, 2004; Hashmi *et al.*, 2011; Joshi *et al.*, 2015; Sachindra *et al.*, 2018a; Sachindra *et al.*, 2018b; Ren *et al.*, 2019; Sachindra and Kanae, 2019). GP-based symbolic regression typically achieves an about 10% smaller RMSE compared to linear regression (e.g., Coulibaly, 2004; Hashmi *et al.*, 2011).

1.2 | Conflicting objectives

Solely minimizing the RMSE provides an estimate of the expected value $E(y|\mathbf{X})$ of the local variable y given the

larger-scale predictors \mathbf{X} . Let us write the time series of the local variable as

$$\mathbf{y} = E(\mathbf{y}|\mathbf{X}) + \boldsymbol{\varepsilon} \quad (1)$$

with $\boldsymbol{\varepsilon}$ (which has mean zero) denoting the error between the actual values of \mathbf{y} and its expectation given \mathbf{X} . If $E(\mathbf{y}|\mathbf{X})$ is modelled by linear regression, then $\boldsymbol{\varepsilon}$ is the component of \mathbf{y} that cannot be linearly described by \mathbf{X} . The series of expected value predictions $\mathbf{y}^{eds} = E(\mathbf{y}|\mathbf{X})$ neglects $\boldsymbol{\varepsilon}$ and has thus by design a lower variance than \mathbf{y} . (The subscript *eds* denotes an expected-value downscaling). As impact models usually require local climate information with realistic variability, most downscaling methods model also $\boldsymbol{\varepsilon}$.

There are two conceptually different approaches to increasing the variance of an estimated local series: the deterministic and the stochastic approach; both are used in this study. The stochastic approach is widespread in the downscaling community and an integral step in stochastic weather generators. WGs typically obtain realizations of a local variable \mathbf{y}^{sds} by sampling from the conditional distribution (The acronym *sds* denotes a stochastic downscaling.). An example of a deterministic technique that increases the variance of an estimated local series is variance inflation, pioneered by Klein *et al.* (1959) in the context of weather forecasting and by Karl *et al.* (1990) in the context of GCM downscaling. Variance inflation in its most basic form applies a constant factor to a series of expected value estimates \mathbf{y}^{eds} in the form

$$\mathbf{y}_t^{ids} = \frac{\sigma(\mathbf{y}^{obs,tr})}{\sigma(\mathbf{y}^{eds,tr})} \left(\mathbf{y}_t^{eds} - \overline{\mathbf{y}^{eds}} \right) + \overline{\mathbf{y}^{eds}}, \quad (2)$$

where $\sigma(\mathbf{y}^{obs,tr})$ denotes the standard deviation (SD) of the observations during the training period, $\sigma(\mathbf{y}^{eds,tr})$ the SD of the expected value series during the training period, \mathbf{y}_t^{eds} the expected value estimate at time t , and $\overline{\mathbf{y}^{eds}}$ the average over the expected value estimates (The acronym *ids* denotes inflated downscaling.).

There has been some controversy on the validity of inflation and related deterministic techniques in downscaling (see von Storch (1999); Maraun (2013) and comments on the latter by Bürger (2014); Maraun (2014); Glahn (2016)) of which we here only recall some major points: It is known that different local observations can be consistent with the same larger-scale atmospheric state, that is, a local variable is in fact not deterministically determined by the larger-scale state (von Storch, 1999). Moreover, inflation-like techniques transfer the spatio-temporal correlation structure from the

larger to the smaller-scale and can affect local trend estimates (Maraun, 2013). As downscaling estimates inevitably contain uncertainty, that is, there will always be some $\boldsymbol{\varepsilon}q$, a stochastic approach is appropriate. But as pointed out by Glahn (2016) some users may require a specific value prediction not a probability density. A fair comparison between deterministic and stochastic approaches for a specific value prediction with realistic variability may thereby rather compare a single realization of a stochastic model and the deterministically inflated series.

Both, a deterministically inflated series and a realization drawn from a stochastic model will have an increased RMSE compared with an expected value downscaling. One may therefore view the aim of a low RMSE AND recovering observed variance—or more general the probability density of the observation—as conflicting objectives. Evolutionary computation offers a variety of different algorithms for multi-objective optimization (e.g., Coello, 2006; Emmerich and Deutz, 2018) which have been applied in diverse areas such as economics and finance (e.g., Tapia and Coello, 2007), mechanical engineering (e.g., Chiandussi *et al.*, 2012), or time scheduling (e.g., Silva *et al.*, 2004). Most algorithms for multi-objective optimization are based on the concept of Pareto optimality and do not return a single solution for a given optimization problem, but a set of Pareto optimal solutions as different achievable trade-offs between the conflicting objectives.

1.3 | Study outline

We use multi-objective genetic programming to generate downscaling models for estimating local precipitation series at six Alpine stations. In this study, each station is modelled individually, that is, independent of the remaining five stations. We use a fivefold cross-validation. The models are trained to simultaneously minimize the RMSE and the difference between the cumulative probability densities of downscaled and observed/reference series. Multi-objective GP does not return a single downscaling model for each station (and cross-validation period), but a Pareto set of downscaling models representing different trade-offs between optimal RMSE and variability.

Multi-objective GP has been set up to generate deterministic and stochastic downscaling models. In the deterministic version, we evolve symbolic regression equations returning downscaled precipitation as a deterministic function of the larger-scale predictors. In the stochastic version, a two-step procedure is carried out: In the first step, precipitation occurrence is estimated with a standard (i.e., not GP-based) logistic regression, and in

the second step, GP is used to estimate precipitation amounts by sampling from a gamma distribution with parameters conditioned on the larger-scale predictors using GP-based symbolic regression.

A reasonable generalization performance of empirical-statistical downscaling requires the stationarity assumption to be sufficiently met and the prevention of overfitting. The stationarity assumption, that is, the assumption of stationarity of a relation between local predictand and larger-scale predictors, is crucial in any empirical-statistical downscaling and especially important for climate change studies (e.g., Frias *et al.*, 2006; Vrac *et al.*, 2007b; Schmith, 2008; Gutiérrez *et al.*, 2013; Hewitson *et al.*, 2014; Dayon *et al.*, 2015; Dixon *et al.*, 2016). Overfitting can be caused by fitting overly complex downscaling models to an insufficiently large training data sample; this is especially relevant for a highly flexible non-linear method like GP. Sachindra *et al.* (2018b) raised particular concerns towards the generalization performance of GP-based downscaling. Hence, we explicitly compare the generalization performance with that of a GLM, which serves as a benchmark.

Each single downscaling model from the stochastic GP setup contains uncertainty estimates of the downscaled precipitation in form of the fitted probability distributions and can hence be viewed and evaluated as an ensemble. The deterministic downscaling models generated by GP do not provide uncertainties. However, since we are provided with a Pareto set of downscaling models, the resulting predictions may be considered as a pseudo-ensemble. As the Pareto sets were not generated to provide distributional estimates and the resulting ensembles were not calibrated, and further because the members are not independent and not expected to be equally probable, we refer to the ensembles from the full Pareto sets as ‘pseudo-ensembles’. We evaluate ensembles generated from selected stochastic downscaling models as well as pseudo-ensembles generated from the full sets of deterministic or stochastic Pareto optimal models using a proper scoring rule.

The deterministic multi-objective GP has originally been implemented for the downscaling of near-surface atmospheric fields at the meso-scale (Zerenner *et al.*, 2016). An earlier version of the deterministic multi-objective GP for the downscaling of local station observations has contributed to the downscaling method intercomparison coordinated by the European Cooperation in Science & Technology (COST) action VALUE, *Validating and Integrating Downscaling Methods for Climate Change Research* (Gutiérrez *et al.*, 2019). First comparisons of deterministic and stochastic downscaling of precipitation time series with multi-objective GP have appeared in the *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (Zerenner *et al.*, 2018).

2 | EXPERIMENT DESIGN AND DATA

We follow the Experiment 1(a) set up by the COST action VALUE for which detailed descriptions and data are available online at http://www.value-cost.eu/validation/#Experiment_1a (last accessed June 2020). Experiment 1(a) uses predictors from the ERA-Interim reanalysis (Dee *et al.*, 2011; Marun *et al.*, 2015). The experiment tests the capability of a downscaling technique to estimate point (station) data from the European Climate Assessment and Data (ECA&D) data set (Klein Tank *et al.*, 2002) for temperature (daily maximum, minimum and mean) and daily accumulated precipitation at 86 European stations. In this study we downscale daily accumulated precipitation for six stations in the Alpine region (Figure 1, Table 1). We picked the Alpine region as it contains mountain stations such as Saentis with high precipitation amounts and strong temporal variability for which downscaling is expected to be challenging as well as stations like Salzburg where GCM precipitation and local observations differ comparably little.

The COST-VALUE experiment provides a standard pre-selection of commonly used predictors (http://www.value-cost.eu/WG2_predictors; last accessed June 2020) of which we have selected a subset by excluding strongly correlated predictors. We have applied some transformations to make the search for downscaling models more efficient (Table 2). For instance, instead of using the temperatures at 500 and 850 hPa, we have used their average and difference, the latter serving as an indicator of atmospheric stability. Instead of the u – and v – components of the wind vector, we have used wind speed and direction (angle). All predictors have been normalized to zero

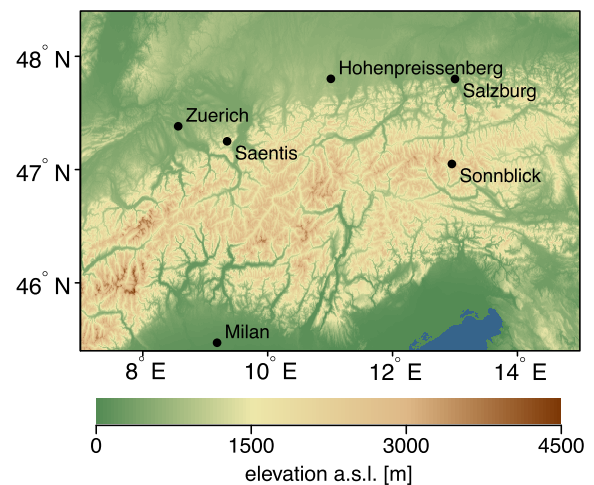


FIGURE 1 Location of the six stations in the alpine region within terrain [Colour figure can be viewed at wileyonlinelibrary.com]

Id	Name	Lon (°E)	Lat (°N)	Altitude (m.a.s.l.)
14	Salzburg	13.0000	47.8000	437
15	Sonnblick	12.9500	47.0500	3,106
48	Hohenpreissenberg	11.0117	47.8017	977
173	Milam	9.1892	45.4717	150
243	Saentis	9.3500	47.2500	2,502
244	Zuerich	8.5667	47.3831	556

TABLE 1 Name, location and altitude of the six stations in the Alpine region

TABLE 2 GCM grid scale predictors available to GP

Variable	Description
X1	\bar{z} Average over 500 and 850 hPa geopotential height $\bar{z} = (z_{500} + z_{850})/2$
X2	z_{1000} 1,000 hPa geopotential height
X3	\bar{T} Average over temperatures in 500 and 850 hPa $\bar{T} = (T_{500} + T_{850})/2$
X4	ΔT Temperature difference between 500 and 800 hPa $\Delta T = T_{500} - T_{850}$
X5	Q_{500} Specific humidity in 500 hPa height
X6	Q_{850} Specific humidity in 850 hPa height
X7	v_{500} Horizontal wind speed in 500 hPa
X8	φ_{500} Wind direction (angle) in 500 hPa
X9	v_{850} Horizontal wind speed in 850 hPa
X10	φ_{850} Wind direction (angle) in 850 hPa
X11	P Daily accumulated precipitation

Note: Predictors have been normalized to zero mean and unit variance, except for P which has been normalized to unit variance only.

mean and unit variance, except for precipitation, which has been normalized to unit variance only.

The experiment covers the time period from January 1, 1979, to December 31, 2008, and is carried out as a five-fold cross-validation by splitting the data set into five sub-periods of 6 years each (1979 to 1984, ..., 2003 to 2008). Each sub-period is successively used as the validation period, while the other 4 sub-periods are used for training.

3 | METHODOLOGY

3.1 | Multi-objective optimization

Multi-objective optimization algorithms address optimization tasks involving multiple conflicting objectives—in the context of downscaling for instance a low RMSE and consistency between the probability distributions of downscaled and observed time series. For optimization problems with conflicting objectives, there is usually no

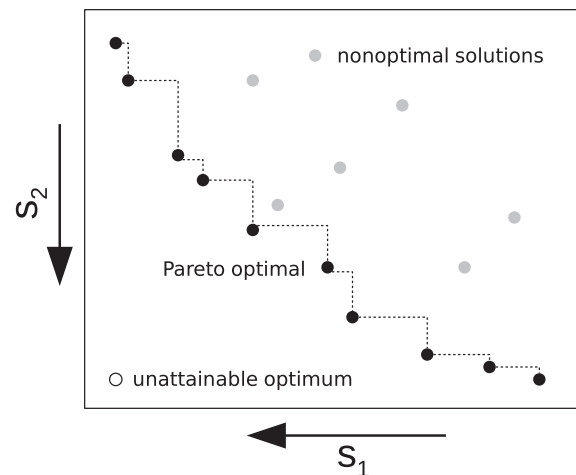


FIGURE 2 The concept of Pareto optimality for two conflicting objectives s_1 and s_2 . The arrows point in the direction of improvement/reduction of error. The theoretical optimum in the lower-left corner is unattainable. The black points are the Pareto optimal solutions, that is, the feasible trade-offs between optimizing w.r.t. s_1 and s_2 . The grey points are the non-optimal solutions, meaning for each grey point there is at least one solution performing favourably for both s_1 and s_2

optimal solution in the absolute sense, but a set of possible compromises between the conflicting objectives. This so-called Pareto set—in the case of two objectives also called Pareto front—contains the Pareto optimal solutions, that is, the solutions for which no other solution exist that is better with respect to all objectives. Figure 2 shows an example of a Pareto front w.r.t. two conflicting objectives s_1 and s_2 . The theoretical optimum solution in the lower left corner is unattainable. The black dots indicate the best achievable compromises, that is, the Pareto set or front. For any pair of Pareto optimal solutions (any two black dots) one performs better w.r.t. s_1 (lower value on the s_1 -axis) and the other performs better w.r.t. s_2 (lower value on the s_2 -axis). All non-optimal solutions (grey dots) are outperformed by at least one Pareto optimal solution.

Simply adding up multiple objectives to a single fitness function usually makes no sense, because the objectives may have different, a priori unknown, ranges, or

even units, and also because a complex fitness function may have multiple local minima, making optimization more difficult. Numerous algorithms deal with multi-objective optimization by means of evolutionary approaches (Emmerich and Deutz, 2018). In this study we use the Strength Pareto Evolutionary Algorithm (SPEA) (Zitzler and Thiele, 1999). In SPEA the fitness of a solution does not depend on the actual values that it achieves for the different objectives, but on a ranking of the solutions of a generation. The fitness of a solution in SPEA is thereby invariant to normalization or scaling of the objectives. SPEA returns not a single optimal solution but a set of Pareto optimal solutions—downscaling models in our application.

3.2 | Deterministic downscaling with multi-objective GP

Genetic programming (GP) belongs to automatic programming, that is, it generates code to solve a given task without the user having to specify the structure of the solution. Instead, the user prescribes the allowed elements of the solution and a fitness function that quantifies how well a suggested solution solves the given task. GP is an evolutionary computation technique; thus, potential solutions for the given task are evolved in analogy to biological evolution. A general introduction to GP can be found for instance in Poli *et al.* (2008).

In tree-based GP the solutions—the downscaling models in our application—are encoded as parse trees (cf. Figure 3). A parse tree is evaluated from the bottom to the top and contains two types of elements also referred to as nodes: functions and terminals. Functions require one or more inputs, that is, one or more branches initiate at each function node. Terminals are zero-argument functions that terminate the tree branches.

Our terminal set comprises the large-scale predictors X_i (cf. Table 2) and numerical constants, which are randomly drawn from a uniform distribution on [0, 1]. By choosing the large-scale predictors for the terminal set, the user performs a pre-selection of predictors. However, how many and which predictors to use in a downscaling model is up to GP and optimized during the evolution. Our function set includes arithmetic functions (addition, multiplication, subtraction), the arctangent (*atan*), and the analytical quotient (AQ) of two variables a and b which is defined as

$$AQ(a,b) = \frac{a}{\sqrt{1+b^2}}. \tag{3}$$

Using AQ instead of a standard or protected division operator avoids the singularity at $b = 0$ and steep gradients in its vicinity. It has been shown that including AQ instead of protected division results in lower errors for various regression tasks (Ni *et al.*, 2012). The parse trees encoding the deterministic downscaling models are allowed to contain 8 levels at maximum.

Figure 3 shows an example of the tree representation of a deterministic downscaling rule. The lowest node of the tree is a numerical constant (0.60752). The constant serves as an input argument to an arcus tangens (*atan*). The *atan* (0.60752) is then multiplied by X_5 and so on. The evaluation of the top node, also called root node—a multiplication in our example—returns the final downscaled precipitation.

As an evolutionary computation technique, GP works with a population of trees which evolves over several generations, and which we limit to 200 trees. The initial generation is randomly generated from the available functions and terminals. Each tree is applied to the given task and evaluated according to the fitness function(s). To generate the next generation the following three steps are repeated until 200 trees have been produced: (a) The operation (crossover or mutation) for generating a new tree is selected. Crossover is selected with a probability of $p_c = 0.8$, mutation with a probability of $p_m = 0.2$. (b) The parent trees are selected from the current generation. The better a tree performs according to the fitness measure(s), the more likely it is to be selected. For crossover two parent trees are required. For mutation only one parent tree is required. (c) The operation is performed on the parent trees. For mutation the tree is cut at a randomly selected

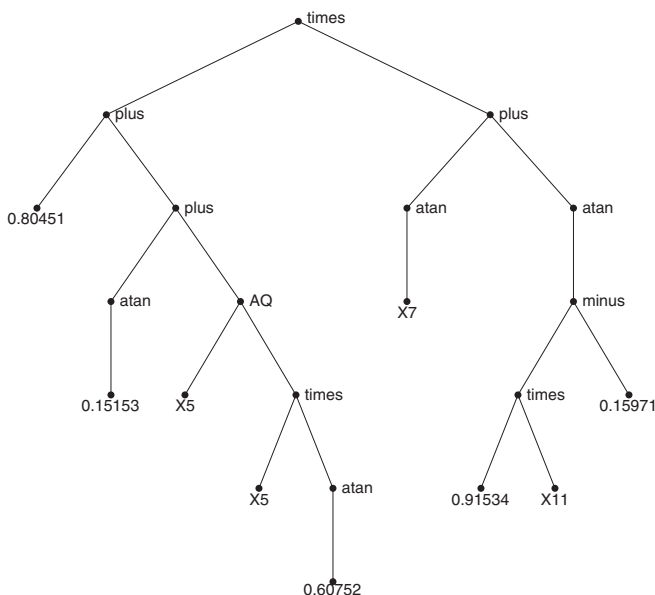


FIGURE 3 Tree representation of a deterministic downscaling model evolved with GP. Shown is one of the downscaling models from the Pareto set for the station Salzburg for cross-validation period 4 (1997–2002)

node and the subtree below that node is replaced by a new randomly generated tree. Each mutation thus generates one tree for the next generation. For crossover the two parent trees are cut at randomly selected nodes and the subtrees below those nodes are exchanged. Each crossover thus generates two trees for the next generation. Evolution stops when the termination criterion—in our case 500 generations—is reached.

Our GP code is based on the *GPLAB* by Silva and Almeida (2003), our setup is summarized in Table 3. As in previous studies employing GP to downscaling (e.g., Coulibaly, 2004; Sachindra *et al.*, 2018a; Sachindra and Kanae, 2019), we use GP to perform a symbolic regression. Symbolic regression is a common real-world application of GP and typically aims at minimizing the root-mean-square error (RMSE) between regression estimates and reference. In this study, we include the difference between empirical cumulative densities of regression estimates and reference as an additional objective.

As outlined in Section 1.2 a low RMSE and consistency between the probability distributions of downscaled and observed series are conflicting objectives and can hence be treated as a multi-objective optimization problem. Our first objective, the RMSE between downscaling estimate and reference is given by

$$\text{RMSE}(y^{ds}, y^{obs}) = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t^{ds} - y_t^{obs})^2} \quad (4)$$

where y_t^{obs} denotes the station observations and y_t^{ds} the downscaled precipitation with t indexing the time.

TABLE 3 Genetic programming setup for deterministic (GP_d) and stochastic (GP_s) precipitation downscaling

Parameter	Value
Function set	Arithmetic functions (+, −, ×), AQ, atan
Terminal set	Predictor variables (Table 2), random numbers ∈ [0, 1]
Stop criterion	Reaching generation 500
Population size	200
Max. Pareto set size	100
Genetic operators	(subtree-)mutation ($p_m = 0.2$), crossover ($p_c = 0.8$)
Max. tree levels	8
Objectives	RMSE, IQD (GP _s additionally RMSE of deterministic subtree only)

Our second objective, the integrated quadratic distance (IQD), measures how well the observed cumulative distributions are restored by the downscaling. The IQD is a proper distance measure between probability distributions (Gneiting and Raftery, 2007; Thorarinsdottir *et al.*, 2013). The IQD between the empirical cumulative distribution functions (CDF) of downscaled F^{ds} and observed precipitation F^{obs} is given by

$$\text{IQD}(y^{ds}, y^{obs}) = \int_{-\infty}^{\infty} (F^{ds}(y) - F^{obs}(y))^2 dy. \quad (5)$$

The downscaled and observed time series have been used as samples to estimate the respective (empirical) CDFs as

$$F^{ds/obs}(y) = \frac{1}{n} \sum_{t=1}^n \mathbf{1}_{y_t^{ds/obs} \leq y} \quad (6)$$

where $\mathbf{1}_u$ is an index function with $\mathbf{1}_u = 1$ if the condition u is true.

With the described setup we run GP for each station and cross-validation period (cf. Section 2). For each station and cross-validation period we obtain a set of up to 100 Pareto optimal downscaling models.

3.3 | Stochastic downscaling with multi-objective GP

The multi-objective GP setup described thus far produces deterministic downscaling models; thus each model produces a deterministic (fixed value) estimate of downscaled precipitation at each time t . To set up multi-objective GP to generate stochastic downscaling models we impose certain constraints on the GP trees. Our setup may be viewed as a hybrid of GP-based symbolic regression and a GLM. In fact GP comes in only in the estimation of precipitation amounts, while the probability of precipitation occurrence is modelled using a standard (i.e., not GP-based) logistic regression, as summarized in Table 4. Logistic regression provides the probability of precipitation occurrence $p(y^{0/1} = 1|X)$ given the larger-scale predictors X from Table 2. To obtain a time series of 0 (no precipitation) and 1 (precipitation) we draw a random number z_t from a uniform distribution on the interval [0,1] for each day t and whenever $p(y^{0/1} = 1|X_t) > z_t$ we set $y_t^{0/1} = 1$, that is, day t is a wet day; otherwise $y_t^{0/1} = 0$, that is, day t is a dry day.

To obtain stochastic estimates of precipitation amounts with GP we prescribe the uppermost function of the trees, the root node, to be a random number

TABLE 4 Generation of the downscaled precipitation series y^{ds} in the stochastic GP setup. Two steps are carried out successively

Step 1: Precipitation yes/no ($y^{0/1}$)	Step 2: Precipitation amount (y^a)
(standard) logistic regression $\rightarrow p$ ($y^{0/1} = 1 X$)	GP $\rightarrow \Gamma(\mu(X), \sigma(X))$
If $p(y^{0/1} = 1 X) > z$ then $y^{0/1} = 1$ else $y^{0/1} = 0$ with $z \sim \text{unif}(0, 1)$	$y^a \sim \Gamma(\mu^{in}, \sigma^{in})$

Note: The logistic regression (fitted using maximum likelihood estimation) provides the probability of precipitation occurrence. Precipitation amount is estimated with GP. Precisely, GP estimates the parameters μ and σ of a gamma distribution Γ as functions of the larger-scale predictors X , that is, $\Gamma(\mu(X), \sigma(X))$. Precipitation amounts y^a are then drawn from $\Gamma(\mu(X), \sigma(X))$. The objectives are computed from the final precipitation estimates $y^{ds} = y^{0/1}y^a$ for evaluation in hindsight as well as during evolution.

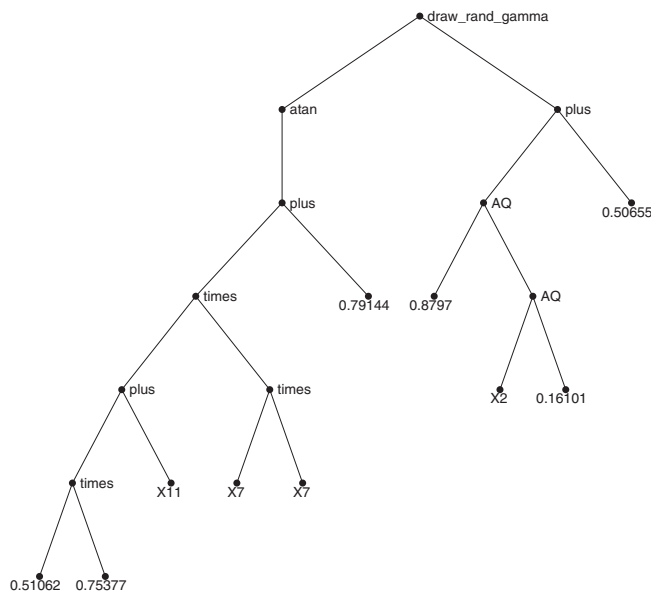


FIGURE 4 Tree representation of a stochastic downscaling model evolved with GP. Shown is one of the downscaling models from the Pareto set for the station Salzburg for cross-validation period 1 (1979–1984)

generator (compare Figure 4). The root node is the last function to be executed in the evaluation of a tree and thus produces the final outcome of the tree. As we are estimating precipitation amounts, we implement a random number generator, that draws a random number y^a from a gamma distribution. The random number generator has two input arguments, the mean μ and the SD σ of the gamma distribution, which are evolved by GP. The available functions and terminals are, except for the random number generator at the root node, the same as in the deterministic GP setup (cf. Table 3). GP may therefore generate negative μ or σ . We deal with this by distinguishing between μ^{out} and σ^{out} returned by the

functions one level below the root node and μ^{in} and σ^{in} , the input to the random number generator at the root node which draws the precipitation amount y^a from the gamma distribution $\Gamma(\mu^{in}, \sigma^{in})$. We set $\sigma^{in} = |\sigma^{out}|$; for $\mu^{out} > 0$ we set $\mu^{in} = \mu^{out}$ and for $\mu^{out} \leq 0$ we directly set the precipitation amount $y^a = 0$. By combining logistic regression and stochastic GP trees we obtain the final precipitation series as $y^{ds} = y^{0/1}y^a$. The downscaled time series y^{ds} are then used to compute the objectives both during evolution, that is, the fitting of the stochastic GP trees, and for the validation later on. In our current implementation a fitness assignment is based on a single realization of y^{ds} .

The described setup offers GP the possibility to produce stochastic models for the precipitation amounts. Whether this possibility is used is up to GP. If it proves advantageous, GP can also develop purely deterministic models by producing $\sigma^{out} = 0$. To reward GP for considering unexplained variability in a stochastic way, we add a third objective in the stochastic GP runs, namely the RMSE of μ^{in} , as

$$\text{RMSE}(\mu^{in}, y^{obs}) = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t^{obs} - \mu_t^{in})^2}. \tag{7}$$

The first two objectives, the RMSE and IQD between downscaling estimate and observations, are the same as in the deterministic setup (Equations (4) and (5)).

Like the deterministic setup, we also run the stochastic GP setup for each station and cross-validation period (cf. Section 2). Thus, we obtain a set of up to 100 stochastic Pareto optimal downscaling models for each station and cross-validation period. It is expected that each Pareto set will contain models with a strong stochastic component (large σ) and a comparatively low IQD as well as models with a weak stochastic component (small or zero σ) that achieve a low RMSE at the expense of a comparatively high IQD.

3.4 | Performance assessment

As a baseline to evaluate performance and generalization performance of the GP-based downscaling, we use a GLM based downscaling approach. Our implementation follows Chandler and Wheeler (2002) who have adopted the two-stage approach of Coe and Stern (1982); Stern and Coe (1984). In stage one precipitation occurrence is modelled with a logistic regression. We use the same large-scale predictors that we have used for the GP-based downscaling. That is, our GLM for precipitation occurrence provides the probability of precipitation on day

t conditioned on the large-scale predictors from Table 2. In stage two precipitation amounts are modelled by fitting gamma distributions to the observed precipitation amounts on wet days only. The mean of the gamma distributions is conditioned on the large-scale predictors from Table 2. The shape parameter is assumed to be constant. To obtain individual time series from our fitted GLMs we proceed analogously to the generation of time series from the stochastic GP models (see Section 3.3).

To assess the performance of GP- and GLM-based downscaling we use the skill scores of the objectives, $\widetilde{\text{RMSE}}$ and $\widetilde{\text{IQD}}$, which we calculate as

$$\widetilde{\text{IQD}} = 1 - \frac{\text{IQD}(y^{ds}, y^{obs})}{\text{IQD}(y^{gcm}, y^{obs})} \quad (8)$$

$$\widetilde{\text{RMSE}} = 1 - \frac{\text{RMSE}(y^{ds}, y^{obs})}{\text{RMSE}(y^{gcm}, y^{obs})}. \quad (9)$$

The precipitation time series of the GCM grid box containing the considered station y^{gcm} serves as our reference. Hence, the skill scores quantify the improvement w.r.t. an objective that is achieved by a downscaling model compared with the raw GCM precipitation. For the six stations (and five cross-validation periods) considered the $\text{RMSE}(y^{gcm}, y^{obs})$ ranges from 5.2 mm (Milan, cross-validation period 5) to 13.8 mm (Saentis, cross-validation period 4). The $\text{IQD}(y^{gcm}, y^{obs})$ ranges from 0.01 mm² (Salzburg, cross-validation period 5) to 0.57 mm² (Saentis, cross-validation period 4). A positive skill score indicates that downscaling obtains an improvement compared with the raw GCM (the larger the skill score the better). Note that the skill scores $\in (-\infty, 1]$. Further note that due to the different magnitudes of the errors (especially for the IQD) of the raw GCM precipitation at distinct stations, one should not use the skill scores for a direct comparison between different stations.

Finally, we study if the GP-based downscaling provides useful uncertainty estimates. We evaluate the performance of ensembles generated from selected stochastic downscaling models as well as pseudo-ensembles generated from the full sets of Pareto optimal models. To do so we use the continuous rank probability score (CRPS), a popular verification tool for ensemble forecasts (Hersbach, 2000; Jordan *et al.*, 2017), which is for a predictive distribution F^{ens} and observations y^{obs} given by

$$\text{CRPS}(F^{ens}, y^{obs}) = \frac{1}{n} \sum_{t=1}^n \int_{y=-\infty}^{y=\infty} \left(F_t^{ens}(y) - \mathbf{1}_{y \geq y_t^{obs}} \right)^2 dy. \quad (10)$$

where $\mathbf{1}$ is a unit step function, that is, $\mathbf{1}_{y \geq y_t^{obs}} = 1$ for $y \geq y_t^{obs}$ and 0 else. The empirical cumulative density F_t^{ens}

TABLE 5 Short description of the four downscaling ensembles

Abbreviation	# members	Description
GP _s	100	100 realizations from a single downscaling model (smallest IQD in training) from the stochastic GP setup
GLM	100	100 realizations from a generalized linear model
GP _s	97–100	Full set of Pareto optimal models from the stochastic GP setup (1 realization from each model)
GP _d	97–100	Full set of Pareto optimal models from the deterministic GP setup

at time t is computed from the ensemble (≈ 100 members, compare Table 5) at the respective point in time y_t^j as

$$F_t^{ens}(y) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{y_t^i \leq y} \quad (11)$$

where i indexes the ensemble members, N denotes the total number of members and index function $\mathbf{1}_u = 1$ if the condition u is true and 0 else. The lower the CRPS, the better. A CRPS of zero is only achieved when all ensemble members match the observations exactly, that is, for a perfect forecast with zero uncertainty.

4 | RESULTS

In the following we first study the performance of the Pareto optimal downscaling models from deterministic and stochastic GP setups w.r.t. the objectives RMSE and IQD. As some studies have reported problems with the generalization ability of GP-based downscaling (e.g., Sachindra *et al.*, 2018b), we focus in particular on the generalization performance. Performance and generalization performance of the GLM-based downscaling serves as a reference. As outlined in Section 1.2 impact models typically require local climate information with realistic variability. We therefore take a closer look at the performance of the downscaling models with the lowest IQD. Finally, we evaluate ensembles generated from selected stochastic downscaling models as well as pseudo-ensembles generated from the full sets of Pareto optimal models to study if the GP-based downscaling provides useful uncertainty estimates.

4.1 | Pareto sets

The Pareto optimal trees returned by the GP algorithm have on average about 50 nodes in the deterministic and about 40 nodes in the stochastic setup. The majority of trees have 8 levels, which is the maximum number of levels allowed in our setups. Figure 3 shows one of the downscaling models for the station Salzburg (cross-validation period 4) from the deterministic GP setup. Figure 4 shows one of the trees for Salzburg (cross-validation period 1) from the stochastic GP setup. The trees shown are comparatively small. The deterministic tree has 21 nodes arranged on 7 levels. The stochastic tree has 20 nodes on 7 levels. Both trees use only 3 out of the 11 predictors offered to GP; both use horizontal wind speed in 500 hPa and daily accumulated precipitation. The deterministic tree further uses specific humidity in 500 hPa as a predictor; the stochastic tree uses the 1,000 hPa geopotential height instead (cf. Table 2).

Figure 5 shows excerpts of the downscaled precipitation time series at Salzburg. Shown are the models with smallest RMSE (6.3 mm; the error values given in this paragraph were obtained from the original, that is, non-normalized, time series.) and the smallest IQD ($1.6 \times 10^{-4} \text{ mm}^2$) in training from the deterministic GP setup (GP_d) together with the observations. We show excerpts from the training period as this allows us to illustrate the systematic difference between the models

irrespective of their generalization performance. As expected, the model with the smallest RMSE underestimates variability in general and extremes in particular and therefore has a comparable high IQD (0.2 mm^2), while the model with the lowest IQD better represents variability at the expense of a higher RMSE (8.7 mm).

Figure 6 shows the full Pareto sets from the deterministic and the stochastic GP setup as well as the GLM for all six stations for validation periods 1 and 5. The area of positive skill is indicated by grey hatching. The deterministic GP provides the expected line-like Pareto front. For most stations and validation periods this shape is largely preserved during validation. The Pareto fronts from the stochastic GP setup are comparably scattered due to the stochastic nature of the models. Note that Figure 6 shows the average skill scores over 100 realizations of a stochastic model, while the fitness assessment during evolution was based on single realizations. Moreover, the stochastic GP setup uses the RMSE of the deterministic subtree as a third objective while Figure 6 shows only two dimensions of the three-dimensional space spanned by the objectives of the stochastic setup.

The Pareto sets in Figure 6 illustrate that RMSE and IQD are indeed conflicting objectives; downscaling models that are optimal w.r.t. the RMSE are suboptimal w.r.t. the IQD and vice versa. Reproducing the temporal variability of the local stations goes along with an increase in RMSE compared to an expected value

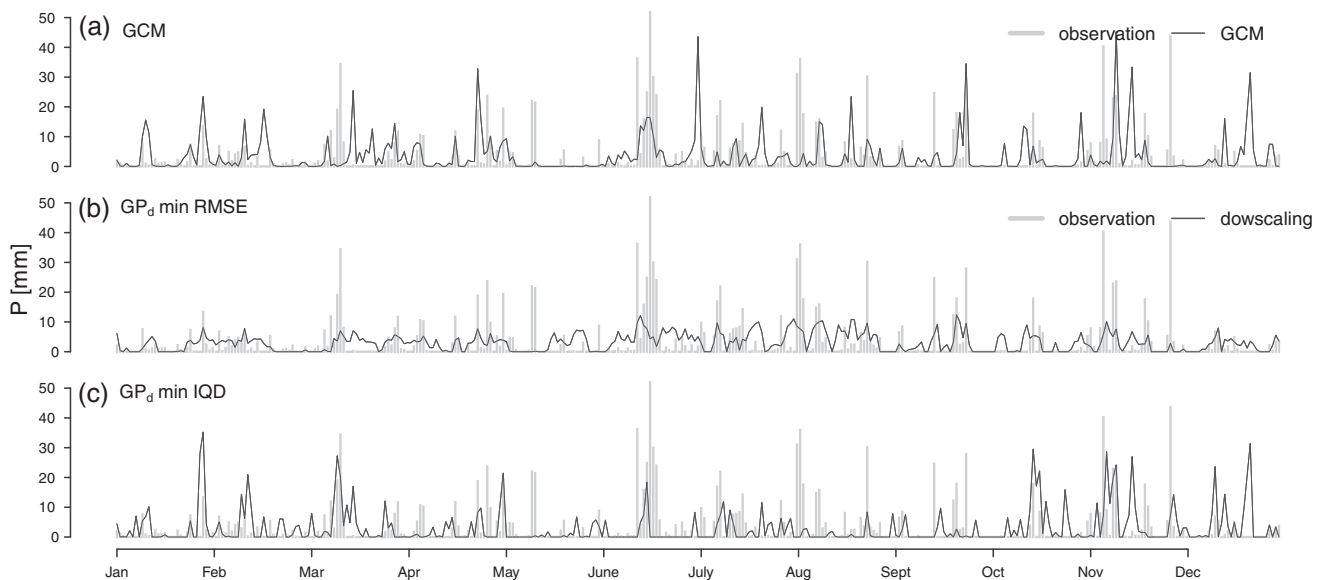


FIGURE 5 Excerpt of the series of daily accumulated precipitation P at Salzburg. Shown is the year 1979. The grey bars show the station observations. The lines show the GCM precipitation at the closest grid point (a) and the downscaled series from the two different downscaling models from the deterministic GP setup (GP_d): The model with the best performance concerning RMSE (b) and the model with the best performance concerning IQD (c). Note that here we show a year from the training period to illustrate the conceptual difference between optimizing for RMSE and IQD irrespective of the generalization performance of the downscaling models

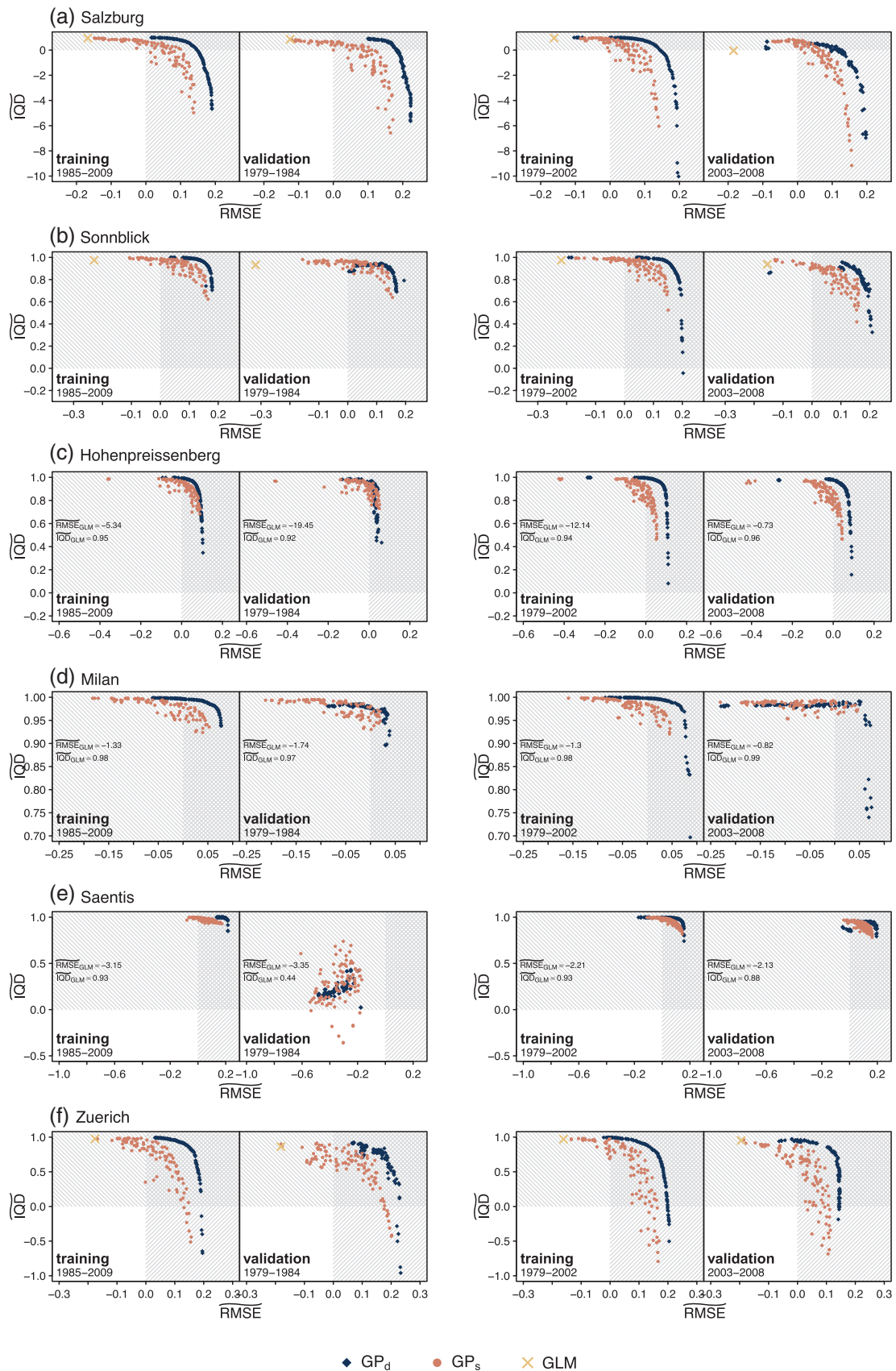


FIGURE 6 IQD skill score (IQD) and RMSE skill score (RMSE) of the Pareto optimal models returned by deterministic GP setup (GP_d), stochastic GP setup (GP_s) and GLM. For stochastic GP and GLM the average IQD and RMSE over 100 realizations of each model are plotted. The plotted regions are adjusted to the ranges of IQD and RMSE for the GP-based downscaling at the respective station. For cases where the GLM lies outside the plotted region, numbers are given in the respective panel [Colour figure can be viewed at wileyonlinelibrary.com]

downscaling. We find the maximum IQD skill to reach values close to 1 for all methods, while the RMSE skill does not exceed 0.25 for any of the methods and cases studied. The deterministic downscaling models with optimum RMSE outperform the stochastic models. Hence, in Figure 6 the Pareto fronts from the deterministic GP setup are shifted to the right compared with the stochastic models, which is most obvious for Salzburg. For the majority of cases deterministic and stochastic models are able to recover variability to a similar extent and perform comparably in terms of IQD. The deterministic models, however, appear to restore variability at a lower RMSE. This is in line with Bürger and Chen (2005) who have shown that (for the training period) restoring variability by randomization in general leads to larger errors than deterministic inflation. However, the stochastic models with optimum IQD outperform the deterministic ones in validation for the majority of cases (most obvious for Milan and Sonnblick). This difference appears relatively small. Nevertheless, no approach is uniquely superior.

The GLM downscaling achieves a similar IQD skill as the GP-based downscaling, but for three stations (Hohenpreissenberg, Milan and Saentis) the realizations drawn from the GLM exhibit much lower RMSE skill. For almost all cases the shapes of the Pareto sets in validation closely resemble the training performance, but with slightly lower skill. Note, that a slightly lower IQD skill for the validation periods is expected as the cumulative densities are estimated from fewer samples (6 years validation compared with 24 years training). The generalization performance of the GP-based downscaling is throughout on par with the generalization performance of the GLM downscaling. For some cases we find differences in generalization performance between the methods: For Sonnblick (validation period 1) the stochastic GP downscaling outperforms the deterministic GP in validation. A similar behaviour, but less pronounced, is observed for Milan. For Saentis (validation period 1) all methods fail; the SD of the observed precipitation is 28% smaller in validation than in training and none of the downscaling methods appropriately reproduces this difference.

4.2 | Selected downscaling models

We now take a closer look at the performance and generalization performance of the downscaling models that best represent variability; that is, from each Pareto set we select the downscaling model with the smallest IQD on the training data set. Figure 7 shows the IQD skill score for the respective deterministic GP model and for 100 realizations of the respective stochastic GP model

together with the GLM. Again, results strongly vary between stations and validation periods. There are cases, Zuerich and Salzburg (except for cross-validation period 5) for instance, where none of the models is clearly superior. For Milan, Hohenpreissenberg and Saentis the GP models tend to achieve a higher IQD skill compared to the GLM; exceptions are validation period 2 for Hohenpreissenberg, validation period 5 for Milan, and validation period 1 for Saentis. For Sonnblick we find a lower IQD skill for the deterministic GP models compared with the stochastic ones for cross-validation periods 1 and 5. Here the deterministic GP models appear to generalize less well than stochastic GP and GLM. A similar situation is found for Hohenpreissenberg for validation period 2 and for Salzburg for validation period 5. As already observed in Figure 6 the most difficult case appears to be Saentis, validation period 1 for which all 3 methods yield a much lower IQD skill in validation compared with training.

The downscaling models performing well w.r.t. the IQD have a higher RMSE than the original (not down-scaled) GCM precipitation (Figure 8). Since the GCM precipitation represents grid box averages its time series typically exhibit lower temporal variability than local station observations and downscaling typically increases variability. The GP models have a higher RMSE skill compared with the GLM because the RMSE serves as an objective in our GP setup. The GP models thus try to recover the observed cumulative density while increasing the RMSE as little as possible. For the majority of cases the deterministic GP yields the highest RMSE skill, followed by the stochastic GP, and the GLM. For the stations Hohenpreissenberg, Milan, and Saentis we find the RMSE skill to vary strongly between the GLM realizations which is caused by the pronounced tail of the fitted gamma distributions for these stations. For the RMSE skill, we again observe differences between training and validation periods, but unlike for the IQD in both positive and negative direction. Overall the differences are the smallest for the stochastic GP models.

4.3 | Ensembles

In the following we evaluate the capability of the GP-based downscaling to provide distributional estimates that capture the stochastic uncertainty. We evaluate the performance using the CRPS (Equation (10)). We evaluate three different GP-based (pseudo-) ensembles. An ensemble generated from the GLM serves as a reference. Figure 9 shows all four ensembles each containing ~ 100 members (see also Table 5) for the station Salzburg. \widehat{GP}_s contains a single realization of each Pareto optimal

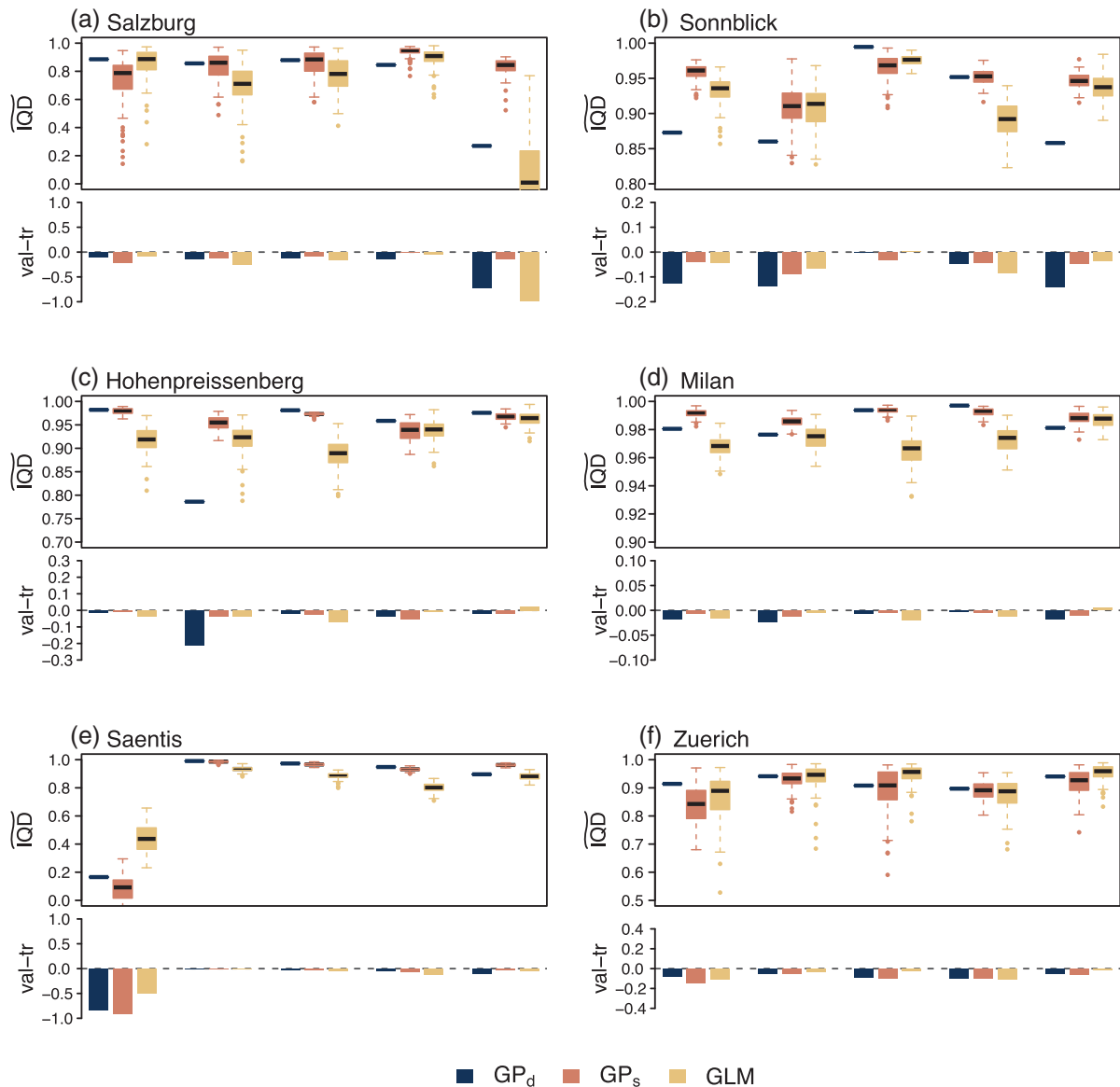


FIGURE 7 IQD skill score (\widehat{IQD} , top) and difference in IQD skill between validation (val) and training (tr) period (bottom) for all five cross-validation periods (left to right) for all six stations (a–f). Shown are one model from the deterministic GP setup (GP_d), one model from the stochastic GP setup (GP_s) and a generalized linear model (GLM). From the Pareto sets returned by GP, we have selected those models that yield the lowest IQD during training. For stochastic GP and the GLM 100 realizations have been drawn for each station/cross-validation period. For the bottom panels we have first computed the average IQD skill over 100 realizations for training and validation separately and then taken the difference of the average skill between the training and validation period [Colour figure can be viewed at wileyonlinelibrary.com]

downscaling model from the stochastic GP setup. GP_s denotes an ensemble based on a single stochastic downscaling model. To obtain the GP_s ensemble we draw 100 realizations from the stochastic GP model with the lowest IQD on the training data set. \widehat{GP}_d contains all Pareto optimal downscaling models from the deterministic GP setup. For comparison we also include a GLM ensemble of 100 realizations. Note that the pseudo-ensembles \widehat{GP}_s and \widehat{GP}_d occasionally contain fewer than 100 members; Typically the Pareto sets reach the

maximum allowed number of 100 members during evolution, but in each generation it can happen that more models are removed than added, for instance when a newly evolved model outperforms two or more members of the Pareto set.

Already from the 1-year excerpts from Salzburg shown in Figure 9 we obtain a first insight into the ensemble performances: \widehat{GP}_d exhibits the smallest spread, followed by GP_s and \widehat{GP}_s with similar spread, and finally the GLM with the largest spread, most obvious for the

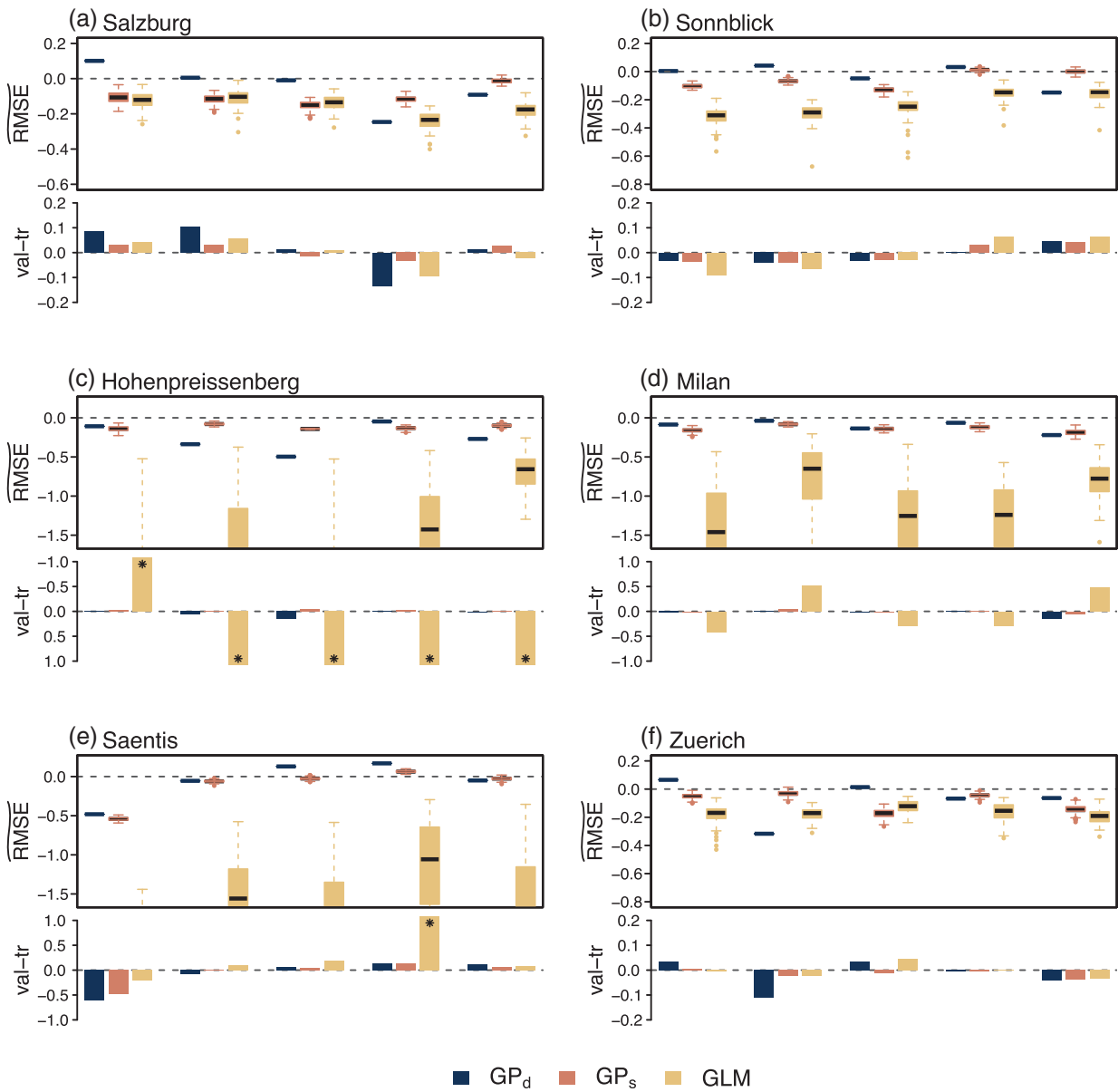


FIGURE 8 RMSE skill score (\overline{RMSE} , top) and difference in RMSE skill between validation (val) and training (tr) period (bottom) for all five cross-validation periods (left to right) for all six stations (a–f). Shown are one model from the deterministic GP setup (GP_d), one model from the stochastic GP setup (GP_s) and a generalized linear model (GLM). From the Pareto sets returned by GP, we have selected those models that yield the lowest IQD during training. For stochastic GP and the GLM 100 realizations have been drawn for each station/cross-validation period. For the bottom panels we have first computed the average RMSE skill over 100 realizations for training and validation separately and then computed the difference of the average skill between training and validation period [Colour figure can be viewed at wileyonlinelibrary.com]

months June to August. Figure 10 shows the CRPS of the four ensembles for all stations and validation periods. The CRPS is highest for the $\overline{GP_d}$ ensemble foremost due to its too small spread caused by ensemble members with low RMSE. The GLM and the two stochastic GP ensembles are comparably close to each other with station-dependent ranking. We find, however, that for GP_s the CRPS is smaller than for $\overline{GP_s}$ for all cases. The larger CRPS for the $\overline{GP_s}$ pseudo-ensemble is, similar to $\overline{GP_d}$,

presumably caused by ensemble members with low RMSE. For the majority of cases the GP_s ensemble achieves a lower CRPS than the GLM ensemble. Hence, the stochastic GP solutions with optimum IQD, although optimized using the two objectives RMSE and IQD, provide an ensemble that performs well also in terms of the CRPS. This is especially obvious for the stations Hohenpreissenberg, Milan and Saentis. For the stations Salzburg, Sonnblick and Zuerich the differences

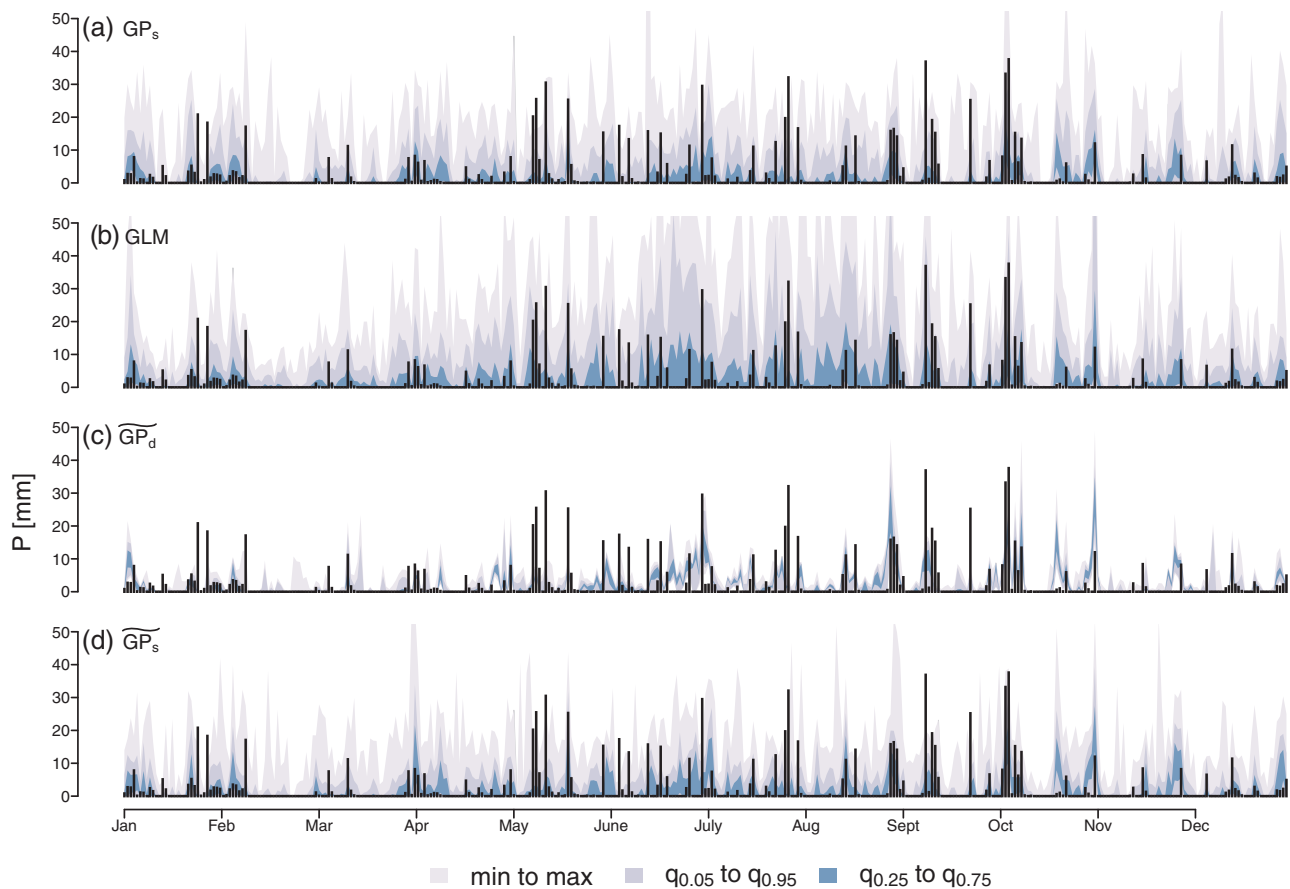


FIGURE 9 Excerpt of the series of daily accumulated precipitation P at Salzburg. Shown in the year 2003. The black bars show the station observations. The shaded areas show the predictions from the four different ensembles of ≈ 100 members each. The area between minimum and maximum over the ≈ 100 members is shaded in the lightest colour, followed by the range between 5- and 95%-quantile and the interquartile range in the darkest colour [Colour figure can be viewed at wileyonlinelibrary.com]

are minor. The differences in CRPS between training and validation period are similar for all four ensembles. The GP-based ensembles thus exhibit a similar generalization performance as the GLM ensemble.

5 | DISCUSSION

We have introduced deterministic and stochastic GP-based approaches for multi-objective precipitation downscaling, which simultaneously minimize the RMSE between observed and downscaled time series and the IQD, which measures the difference between the probability densities of downscaled and observed time series. The Pareto optimal downscaling models in terms of RMSE and IQD show how a low RMSE and a realistic variability are indeed conflicting objectives. As outlined in the introduction it is widely known that restoring variability in regression-based downscaling, either deterministically or by drawing a realization from a stochastic model, increases the RMSE compared with an expected

value downscaling. Our multi-objective downscaling provides an additional approach to restoring variability by minimizing the IQD while keeping the RMSE as small as possible. The multi-objective optimization finds Pareto-optimal solutions within the range from (purely deterministic) expected value downscaling to a fully stochastic downscaling.

The stochastic GP-generated downscaling models with lowest IQD, perform slightly better than the GLM in terms of IQD (Figure 7). The GP-generated downscaling models further achieve a smaller RMSE than the realizations drawn from the GLM (Figures 6 and 8). In our study the generalization performance of both deterministic and stochastic GP-based downscaling is on par with the GLM-based downscaling (Figures 6,7,8). The stochastic solutions with optimum IQD, while optimized using the two objectives RMSE and IQD, provide an ensemble that performs well in terms of the CRPS. The GLM and a single appropriately selected stochastic GP-based model achieve a comparable CRPS (Figure 10), with the stochastic GP model being slightly better for most stations. The

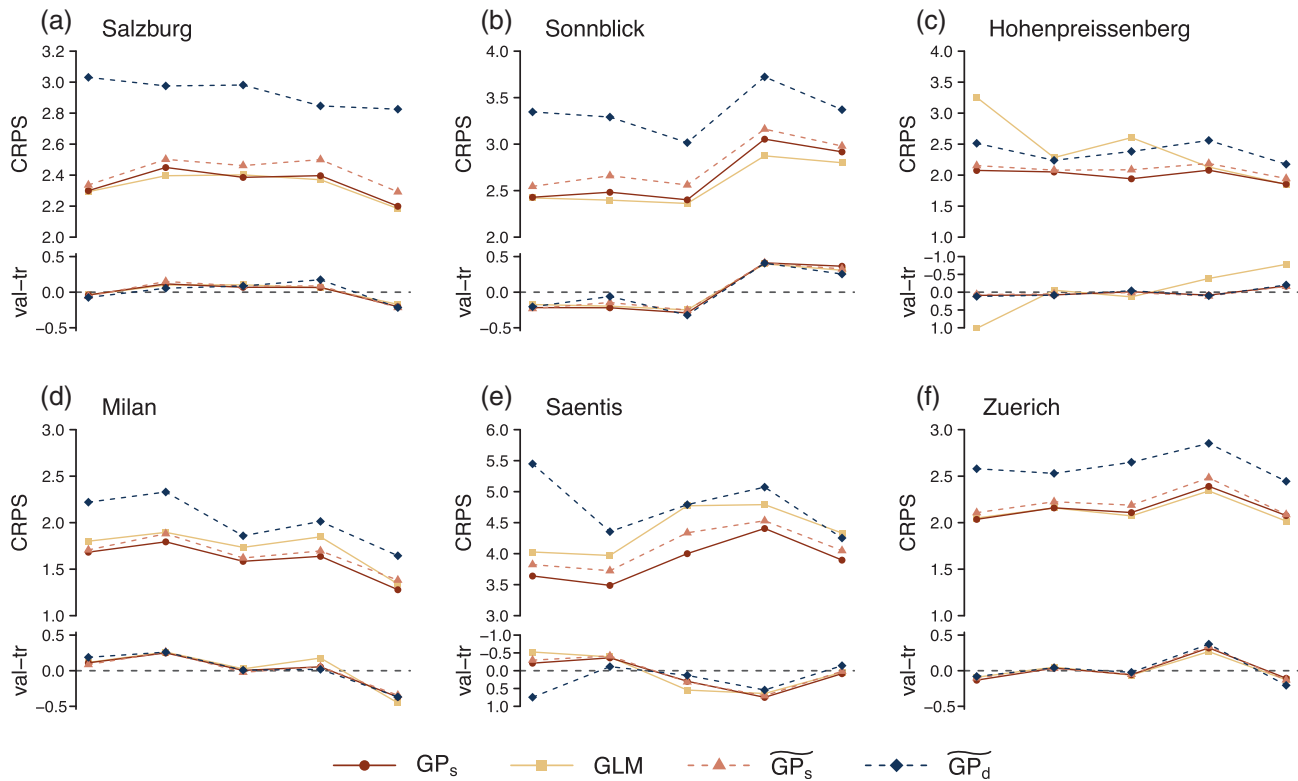


FIGURE 10 CRPS of the four different ensembles for all five cross-validation periods (left to right) and for all six stations (a–f) [Colour figure can be viewed at wileyonlinelibrary.com]

spread of the pseudo-ensembles derived from the variety of solutions within the Pareto set does not appropriately represent the stochastic uncertainty. This is due to the nature of the Pareto set, which contains solutions that are distinct with respect to the objectives and are therefore not realizations from the same population.

In the stochastic GP setup one may in the future rather use a probabilistic score such as the CRPS as objective. We expect a stochastic GP-based downscaling trained to optimize the CRPS to exhibit comparable (or slightly better) performance in terms of CRPS than the stochastic downscaling models trained to minimize the IQD under the constraint of keeping the RMSE as low as possible. However, such a change in objectives would forbid a direct comparison between deterministic and stochastic GP setups. One may further consider the decomposition of the CRPS into reliability and resolution (Hersbach, 2000) for a multi-objective optimization of stochastic downscaling models. The reproduction of temporal (auto-)correlation, and spatial (inter-station) correlation are also considered important for many downscaling products (Maraun *et al.*, 2019; Widmann *et al.*, 2019); such measures could be included easily in a multi-objective GP-based downscaling.

Algorithmic diversity as well as predictor diversity is beneficial in downscaling. Also, in this study it depends

on the station and on the evaluation criterion which downscaling method and which predictors perform favourable. Intercomparisons of downscaling techniques seldom identify a single technique as best (Frost *et al.*, 2011; Gutmann *et al.*, 2014; San-Martín *et al.*, 2017; Gutiérrez *et al.*, 2019; Hertig *et al.*, 2019; Maraun *et al.*, 2019; Soares *et al.*, 2019; Widmann *et al.*, 2019). The choice of a downscaling method is best guided by the requirements of a particular application. Especially when comparing many different methods, the concept of Pareto optimality may be useful. Given a set of user-selected performance criteria, the methods can be reduced to a subset of Pareto optimal methods to choose from. The Pareto front or projections of the higher-dimensional Pareto plane may help to visualize trade-offs and contribute to a better informed choice of a particular downscaling method.

We have demonstrated that GP is a flexible technique and not restricted to regression like applications solely producing expected value estimates. The stochastic GP downscaling models provide useful distributional estimates (Figure 10). In the current stochastic GP setup we have assumed precipitation amounts to follow a gamma distribution; thus our GP setup may be viewed as a cross-over between a vectorized GLM and GP-based symbolic regression. Further extensions (implementing

appropriate functions and constraints on the tree structures) may allow for a selection or combination of different probability distributions, similar to mixture models. In addition, downscaling models containing both deterministic and stochastic components could be realized. Further studies combining established downscaling techniques with evolutionary optimization can be found in Horton *et al.* (2017, 2018); Horton (2019) who optimize an analog method using genetic algorithms.

An example of a study considering distributional downscaling by means of a neural network can be found in Carreau and Vrac (2011), who utilized neural networks to fit conditional mixture models for precipitation downscaling. In a recent study by Shi (2020) convolutional neural networks have been trained to partition given data into extreme and non-extreme precipitation allowing for an explicit downscaling of extreme precipitation events. Also in ensemble post-processing evolutionary and neural network based approaches have gained attention (e.g., Bakhshaii and Stull, 2009; Roebber, 2015; Dufek *et al.*, 2017; Rasp and Lerch, 2018; Bremnes, 2020; Taillardat and Mestre, 2020; Grönquist *et al.*, 2021).

Stationarity of the relations between local predictands and the larger-scale atmospheric predictors is a basic assumption of any empirical-statistical downscaling and particularly important for obtaining a satisfactory generalization performance under changing climate (Frias *et al.*, 2006; Vrac *et al.*, 2007b; Schmith, 2008; Gutiérrez *et al.*, 2013; Hewitson *et al.*, 2014; Dayon *et al.*, 2015; Dixon *et al.*, 2016). A recent study by Sachindra *et al.* (2018b) raised particular concerns towards the generalization performance of GP-based downscaling. Sachindra *et al.* (2018b) argue that the generalization performance of GP-based downscaling suffers from GP not identifying a unique set of optimal predictors. When rerunning GP several times one will typically obtain different downscaling models using different predictors as GP in itself is a stochastic technique. The user only prescribes the potential predictors, that is, the predictors available to GP, but not how many and which of these to use in a downscaling model. While we fully agree that highly flexible and potentially nonlinear methods like GP need to be treated with caution, we do not see the non-unique predictor selection as problematic in itself. The GP-based downscaling models represent empirical-statistical relations between predictors and predictands which arise from the physics and dynamics of the atmospheric circulation, but cannot itself be viewed as physical relations. The predictors representing different aspects of atmospheric circulation are not independent; thus predictors do—to some extent—contain common information. Methods such as linear regression identify a single optimal set of predictors but there may exist other predictor combinations which do not perform significantly inferior and only infinite sampling

would allow one to be sure of detecting a true unique best predictor set.

We believe the limited generalization performance of GP-based downscaling reported in Sachindra *et al.* (2018b) might not be caused by a non-optimal predictor selection. First, extreme conditions such as drought periods are hard for any empirical-statistical downscaling method, especially when not included in the training period. We observed a similar case for the station Saentis (validation period 1), where both GP and GLM struggle. Second, the use of standard or protected division a/b in symbolic regression is problematic due to the singularity at $b = 0$ and the steep gradient in its vicinity. Replacing the division operator by an analytical quotient has been shown to improve the generalization performance for a wide range of regression tasks (Ni *et al.*, 2012). Hence, a rather conservative function set, excluding exponential function and logarithm, and replacing division operators by an analytical quotient can be beneficial for the generalization performance of GP-based downscaling and may reduce or even prevent GP models from producing unrealistic large outliers as reported in Sachindra *et al.* (2018b) and Sachindra and Kanae (2019).

Even though there is likely not a single, uniquely identifiable optimal predictor set as discussed above, an appropriate predictor selection is crucial for downscaling models, which are robust under changing climate (Schmith, 2008; Gutiérrez *et al.*, 2013). In the present study predictor pre-selection has gained comparably little attention, as our main focus has been the intercomparison of the different downscaling approaches (deterministic and stochastic GP and a GLM). Especially moving away from solely offering grid-point predictors to the GP algorithm, but incorporating principal components of the atmospheric circulation or weather classes may further improve the downscaling performance. Also the use of state-of-the-art reanalysis such as the ERA5 (Hersbach *et al.*, 2020) might further improve downscaling performance. The use of predictors from ERA-Interim in the present study allows a comparison with the results of the COST-VALUE intercomparison studies. Gutiérrez *et al.* (2019) in particular contains an in depth evaluation (e.g., w.r.t. wet-day frequency and seasonality) of a previous version of deterministic multi-objective GP downscaling among and in comparison to a variety of other empirical-statistical downscaling technique. Hertig *et al.* (2019) provides an in-depth evaluation w.r.t. extremes.

ACKNOWLEDGEMENTS

We would like thank the two anonymous referees for their insightful comments and suggestions. The authors further acknowledge the COST Action VALUE on


“Validating and Integrating Downscaling Methods for Climate Change Research” for providing experiment design and daily station data (www.value-cost.eu; station data were extracted from ECA&D public archive, www.ecad.eu/dailydata/). The ERA-Interim reanalysis is publicly available at www.ecmwf.int/en/research/climate-reanalysis/era-interim. The GP code is based on the GPLAB-A Genetic Programming toolbox for Matlab (Silva and Almeida, 2003). For the GLM-based reference methods and the analysis of the downscaling results we used the implementations available in R. This work has been carried out within the Transregional Collaborative Research Center 32 on “Patterns in Soil-Vegetation-Atmosphere-Systems” (Simmer *et al.*, 2015) funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).


Open access funding enabled and organized by Projekt DEAL.

ORCID

Tanja Zerenner  <https://orcid.org/0000-0002-4672-5619>

Victor Venema  <https://orcid.org/0000-0001-6902-4869>

Petra Friederichs  <https://orcid.org/0000-0003-4566-572X>

Clemens Simmer  <https://orcid.org/0000-0003-3001-8642>

REFERENCES

- Abaurrea, J. and Asin, J. (2005) Forecasting local daily precipitation patterns in a climate change scenario. *Climate Research*, 28(3), 183–197.
- Bakhshaii, A. and Stull, R. (2009) Deterministic ensemble forecasts using gene-expression programming. *Weather and Forecasting*, 24(5), 1431–1451.
- Baño-Medina, J., Manzanar, R. and Gutiérrez, J.M. (2020) Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, 13(4), 2109–2124.
- Banzhaf, W., Nordin, P., Keller, R. and Francone, F. (1997) *Genetic Programming: An Introduction: On the Automatic Evolution of Computer Programs and its Applications (the Morgan Kaufmann Series in Artificial Intelligence)*. San Francisco, CA, USA: Morgan Kaufmann Publishers.
- Benestad, R.E., Hanssen-Bauer, I. and Chen, D. (2008) *Empirical-Statistical Downscaling*. Singapore: World Scientific.
- Bremnes, J.B. (2020) Ensemble postprocessing using quantile function regression based on neural networks and bernstein polynomials. *Monthly Weather Review*, 148(1), 403–414.
- Bürger, G. (2014) Comment on “bias correction, quantile mapping, and downscaling: revisiting the inflation issue”. *Journal of Climate*, 27(4), 1819–1820.
- Bürger, G. and Chen, Y. (2005) Regression-based downscaling of spatial variability for hydrologic applications. *Journal of Hydrology*, 311(1–4), 299–317.
- Carreau, J. and Vrac, M. (2011) Stochastic downscaling of precipitation with neural network conditional mixture models. *Water Resources Research*, 47, W10502.
- Chandler, R.E. and Wheater, H.S. (2002) Analysis of rainfall variability using generalized linear models: a case study from the west of Ireland. *Water Resources Research*, 38(10), 1192.
- Cheng, C.S., Li, G., Li, Q. and Auld, H. (2011) A synoptic weather-typing approach to project future daily rainfall and extremes at local scale in Ontario, Canada. *Journal of Climate*, 24(14), 3667–3685.
- Chiandussi, G., Codegone, M., Ferrero, S. and Varesio, F.E. (2012) Comparison of multi-objective optimization methodologies for engineering applications. *Computers & Mathematics with Applications*, 63(5), 912–942.
- Coe, R. and Stern, R. (1982) Fitting models to daily rainfall data. *Journal of Applied Meteorology*, 21(7), 1024–1031.
- Coello, C.C. (2006) Evolutionary multi-objective optimization: a historical view of the field. *IEEE Computational Intelligence Magazine*, 1(1), 28–36.
- Coulibaly, P. (2004) Downscaling daily extreme temperatures with genetic programming. *Geophysical Research Letters*, 31, L16203.
- Coulibaly, P., Dibikey, Y.B. and Ancil, F. (2005) Downscaling precipitation and temperature with temporal neural networks. *Journal of Hydrometeorology*, 6(4), 483–496.
- Dayon, G., Boé, J. and Martin, E. (2015) Transferability in the future climate of a statistical downscaling method for precipitation in France. *Journal of Geophysical Research-Atmospheres*, 120(3), 1023–1043.
- Dee, D., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E. V., Isaksen, I., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.J., Park, B.K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.N. and Vitart, F. (2011) The ERA-interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656), 553–597.
- Dixon, K.W., Lanzante, J.R., Nath, M.J., Hayhoe, K., Stoner, A., Radhakrishnan, A., Balaji, V. and Gaitán, C.F. (2016) Evaluating the stationarity assumption in statistically downscaled climate projections: is past performance an indicator of future results? *Climatic Change*, 135(3–4), 395–408.
- Dufek, A.S., Augusto, D.A., Dias, P.L. and Barbosa, H.J. (2017) Application of evolutionary computation on ensemble forecast of quantitative precipitation. *Computers & Geosciences*, 106, 139–149.
- Emmerich, M.T. and Deutz, A.H. (2018) A tutorial on multi-objective optimization: fundamentals and evolutionary methods. *Natural Computing*, 17(3), 585–609.
- Frias, M., Zorita, E., Fernández, J. and Rodríguez-Puebla, C. (2006) Testing statistical downscaling methods in simulated climates. *Geophysical Research Letters*, 33(19), L19807.
- Frost, A.J., Charles, S.P., Timbal, B., Chiew, F.H., Mehrotra, R., Nguyen, K.C., Chandler, R.E., McGregor, J.L., Fu, G., Kirono, D.G., Fernandez, E. and Kent, D.M. (2011) A comparison of multi-site daily rainfall downscaling techniques under Australian conditions. *Journal of Hydrology*, 408(1–2), 1–18.
- Glahn, B. (2016) Comment on “bias correction, quantile mapping, and downscaling: revisiting the inflation issue”. *Journal of Climate*, 29(23), 8665–8667.

- Gneiting, T. and Raftery, A.E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.
- Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S. and Hoefler, T. (2021) Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A*, 379(2194), 20200092.
- Gutiérrez, J.M., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., Roessler, O., Wibig, J., Wilcke, R., Kotlarski, S., San Martín, D., Herrera, S., Bedia, J., Casanueva, A., Manzanar, R., Iturbide, M., Vrac, M., Dubrovsky, M., Ribalaygua, J., Pórtoles, J., Rätty, O., Räisänen, J., Hingray, B., Raynaud, D., Casado, M.J., Ramos, P., Zerenner, T., Turco, M., Bosshard, T., Štěpánek, P., Bartholy, J., Pongracz, R., Keller, D. E., Fischer, A.M., Cardoso, R.M., Soares, P.M.M., Czernecki, B. and Pagé, C. (2019) 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(9), 3750–3785.
- Gutiérrez, J.M., San-Martín, D., Brands, S., Manzanar, R. and Herrera, S. (2013) Reassessing statistical downscaling techniques for their robust application under climate change conditions. *Journal of Climate*, 26(1), 171–188.
- Gutmann, E., Pruitt, T., Clark, M.P., Brekke, L., Arnold, J.R., Raff, D. A. and Rasmussen, R.M. (2014) An intercomparison of statistical downscaling methods used for water resource assessments in the United States. *Water Resources Research*, 50(9), 7167–7186.
- Hashmi, M.Z., Shamseldin, A.Y. and Melville, B.W. (2011) Statistical downscaling of watershed precipitation using gene expression programming (GEP). *Environmental Modelling and Software*, 26(12), 1639–1646.
- Hersbach, H. (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5), 559–570.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. and Thépaut, J. N. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049.
- Hertig, E., Maraun, D., Bartholy, J., Pongracz, R., Vrac, M., Mares, I., Gutiérrez, J.M., Wibig, J., Casanueva, A. and Soares, P.M. (2019) Comparison of statistical downscaling methods with respect to extreme events over Europe: validation results from the perfect predictor experiment of the cost action value. *International Journal of Climatology*, 39(9), 3846–3867.
- Hewitson, B., Daron, J., Crane, R.G., Zermoglio, M. and Jack, C. (2014) Interrogating empirical-statistical downscaling. *Climatic Change*, 122(4), 539–554.
- Hewitson, B.C. and Crane, R.G. (1996) Climate downscaling: techniques and application. *Climate Research*, 7(2), 85–95.
- Horton, P. (2019) AtmoSwing: analog technique model for statistical weather forecasting and downscaling (v2. 1.0). *Geoscientific Model Development*, 12(7), 2915–2940.
- Horton, P., Jaboyedoff, M. and Obled, C. (2017) Global optimization of an analog method by means of genetic algorithms. *Monthly Weather Review*, 145(4), 1275–1294.
- Horton, P., Jaboyedoff, M. and Obled, C. (2018) Using genetic algorithms to optimize the analogue method for precipitation prediction in the Swiss Alps. *Journal of Hydrology*, 556, 1220–1231.
- Huth, R. (2002) Statistical downscaling of daily temperature in Central Europe. *Journal of Climate*, 15(13), 1731–1742.
- Huth, R., Mikšovský, J., Štěpánek, P., Belda, M., Farda, A., Chládková, Z. and Pišoft, P. (2015) Comparative validation of statistical and dynamical downscaling models on a dense grid in Central Europe: temperature. *Theoretical and Applied Climatology*, 120(3–4), 533–553.
- Jordan, A., Krüger, F. and Lerch, S. (2017), Evaluating probabilistic forecasts with scoringrules, *arXiv preprint arXiv:1709.04743*.
- Joshi, D., St-Hilaire, A., Ouarda, T. and Daigle, A. (2015) Statistical downscaling of precipitation and temperature using sparse bayesian learning, multiple linear regression and genetic programming frameworks. *Canadian Water Resources Journal/Revue Canadienne Des Ressources Hydriques*, 40(4), 392–408.
- Karl, T.R., Wang, W.-C., Schlesinger, M.E., Knight, R.W. and Portman, D. (1990) A method of relating general circulation model simulated climate to the observed local climate. Part i: seasonal statistics. *Journal of Climate*, 3(10), 1053–1079.
- Keller, D.E., Fischer, A.M., Liniger, M.A., Appenzeller, C. and Knutti, R. (2017) Testing a weather generator for downscaling climate change projections over Switzerland. *International Journal of Climatology*, 37(2), 928–942.
- Kilsby, C., Jones, P., Burton, A., Ford, A., Fowler, H., Harpham, C., James, P., Smith, A. and Wilby, R. (2007) A daily weather generator for use in climate change studies. *Environmental Modelling and Software*, 22(12), 1705–1719.
- Klein Tank, A., Wijngaard, J.B., Können, G.P., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., Heino, R., Bessemoulin, P., Müller-Westermeier, G., Tzanakou, M., Szalai, S., Pálsdóttir, T., Fitzgerald, D., Rubin, S., Capaldo, M., Maugeri, M., Leitass, A., Bukantis, A., Aberfeld, R., van Engelen, A.F.V., Forland, E., Miletus, M., Coelho, F., Mares, C., Razuvaev, V., Nieplova, E., Cegnar, T., Antonio López, J., Dahlström, B., Moberg, A., Kirchhofer, W., Ceylan, A., Pachaliuk, O., Alexander, L.V. and Petrovic, P. (2002) Daily dataset of the 20th-century surface air temperature and precipitation for the European climate assessment. *International Journal of Climatology*, 22(12), 1441–1453.
- Klein, W.H., Lewis, B.M. and Enger, I. (1959) Objective prediction of five-day mean temperatures during winter. *Journal of Meteorology*, 16(6), 672–682.
- Koza, J. (1992) *Genetic Programming, on the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press.
- Maraun, D. (2013) Bias correction, quantile mapping, and downscaling: revisiting the inflation issue. *Journal of Climate*, 26(6), 2137–2143.
- Maraun, D. (2014) ‘Reply to “comment on ‘bias correction, quantile mapping, and downscaling: revisiting the inflation issue”’. *Journal of Climate*, 27(4), 1821–1825.
- Maraun, D., Huth, R., Gutiérrez, J.M., Martín, D.S., Dubrovsky, M., Fischer, A., Hertig, E., Soares, P.M., Bartholy, J., Pongrácz, R., Widmann, M., Casado, M.J., Ramos, P. and Bedia, J. (2019) The

- value perfect predictor experiment: evaluation of temporal variability. *International Journal of Climatology*, 39(9), 3786–3818.
- Maraun, D., Wetterhall, F., Ireson, A.M., Chandler, R.E., Kendon, E.J., Widmann, M., Brienen, S., Rust, H.W., Sauter, T., Themeßl, M., Chun, K.P., Goodess, C.M., Jones, R.G., Onof, C., Vrac, M., Thiele-Eich, I. and Thiele-Eich, I. (2010) Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48(3), RG3003.
- Maraun, D. and Widmann, M. (2018) *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge: Cambridge University Press.
- Maraun, D., Widmann, M., Gutiérrez, J.M., Kotlarski, S., Chandler, R.E., Hertig, E., Wibig, J., Huth, R. and Wilcke, R.A. (2015) VALUE: a framework to validate downscaling approaches for climate change studies. *Earth's Future*, 3(1), 1–14.
- Ni, J., Driberg, R.H. and Rockett, P.I. (2012) The use of an analytic quotient operator in genetic programming. *IEEE Transactions on Evolutionary Computation*, 17(1), 146–152.
- Poli, R., Langdon, W. B., McPhee, N. F. and Koza, J. R. (2008), 'A field guide to genetic programming', <http://www.gp-field-guide.org.uk>.
- Rasp, S. and Lerch, S. (2018) Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146(11), 3885–3900.
- Ren, M., Pang, B., Xu, Z., Yue, J. and Zhang, R. (2019) Downscaling of daily extreme temperatures in the yarlung zangbo river basin using machine learning techniques. *Theoretical and Applied Climatology*, 136(3–4), 1275–1288.
- Richardson, C.W. (1981) Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17(1), 182–190.
- Roebber, P.J. (2015) Evolving ensembles. *Monthly Weather Review*, 143(2), 471–490.
- Sachindra, D., Ahmed, K., Rashid, M.M., Shahid, S. and Perera, B. (2018a) Statistical downscaling of precipitation using machine learning techniques. *Atmospheric Research*, 212, 240–258.
- Sachindra, D., Ahmed, K., Shahid, S. and Perera, B. (2018b) Cautionary note on the use of genetic programming in statistical downscaling. *International Journal of Climatology*, 38(8), 3449–3465.
- Sachindra, D. and Kanae, S. (2019) Machine learning for downscaling: the use of parallel multiple populations in genetic programming. *Stochastic Environmental Research and Risk Assessment*, 33(8–9), 1497–1533.
- San-Martín, D., Manzanar, R., Brands, S., Herrera, S. and Gutiérrez, J.M. (2017) Reassessing model uncertainty for regional projections of precipitation with an ensemble of statistical downscaling methods. *Journal of Climate*, 30(1), 203–223.
- Schmith, T. (2008) Stationarity of regression relationships: application to empirical downscaling. *Journal of Climate*, 21(17), 4529–4537.
- Schoof, J.T. and Pryor, S. (2001) Downscaling temperature and precipitation: a comparison of regression-based methods and artificial neural networks. *International Journal of Climatology*, 21(7), 773–790.
- Shi, X. (2020) Enabling smart dynamical downscaling of extreme precipitation events with machine learning. *Geophysical Research Letters*, 47(19), e2020GL090309.
- Silva, J.D.L., Burke, E.K. and Petrovic, S. (2004) An introduction to multiobjective metaheuristics for scheduling and timetabling. In: *Metaheuristics for Multiobjective Optimisation*. Berlin, Heidelberg: Springer, pp. 91–129.
- Silva, S. and Almeida, J. (2003) GPLAB—a genetic programming toolbox for MATLAB. In: *Proceedings of the Nordic MATLAB Conference, Copenhagen, Denmark, 2003*. Citeseer, pp. 273–278.
- Simmer, C., Thiele-Eich, I., Masbou, M., Amelung, W., Bogena, H., Crewell, S., Diekkrüger, B., Ewert, F., Hendricks Franssen, H. J., Huisman, J.A., Kemna, A., Klitzsch, N., Kollet, S., Langensiepen, M., Löhnert, U., Rahman, A.S.M.M., Rascher, U., Schneider, K., Schween, J., Shao, Y., Shrestha, P., Stiebler, M., Sulis, M., Vanderborght, J., Vereecken, H., van der Kruk, J., Waldhoff, G. and Zerenner, T. (2015) Monitoring and modeling the terrestrial system from pores to catchments—the transregional collaborative research center on patterns in the soil-vegetation-atmosphere system. *Bulletin of the American Meteorological Society*, 96(10), 1765–1787.
- Soares, P., Maraun, D., Brands, S., Jury, M.W., Gutiérrez, J.M., San-Martín, D., Hertig, E., Huth, R., Belušić Vozila, A., Cardoso, R. M., Kotlarski, S., Drobinski, P. and Obermann-Hellhund, A. (2019) Process-based evaluation of the value perfect predictor experiment of statistical downscaling methods. *International Journal of Climatology*, 39(9), 3868–3893.
- Stern, R. and Coe, R. (1984) 'A model fitting analysis of daily rainfall data', *Journal of the Royal Statistical Society. Series A (General)*, 147(1), 1–34.
- Taillardat, M. and Mestre, O. (2020) From research to applications—examples of operational ensemble post-processing in France using machine learning. *Nonlinear Processes in Geophysics*, 27(2), 329–347.
- Tapia, M.G.C. and Coello, C.A.C. (2007) Applications of multi-objective evolutionary algorithms in economics and finance: a survey. In: *Proceedings of the IEEE Congress on Evolutionary Computation, Singapore*. IEEE, pp. 532–539.
- Thorarinsdottir, T.L., Gneiting, T. and Gissibl, N. (2013) Using proper divergence functions to evaluate climate models. *Journal of Uncertainty Quantification*, 1(1), 522–534.
- von Storch, H. (1999) On the use of inflation in statistical downscaling. *Journal of Climate*, 12(12), 3505–3506.
- Vrac, M., Stein, M. and Hayhoe, K. (2007a) Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing. *Climate Research*, 34(3), 169–184.
- Vrac, M., Stein, M., Hayhoe, K. and Liang, X.-Z. (2007b) A general method for validating statistical downscaling methods under future climate change. *Geophysical Research Letters*, 34(18), L18701.
- Widmann, M., Bedia, J., Gutierrez, J.M., Bosshard, T., Hertig, E., Maraun, D., Casado, M.J., Ramos, P., Cardoso, R.M., Soares, P. M., Ribalaygua, J., Pagé, C., Fischer, A.M., Herrera, S. and Huth, R. (2019) Validation of spatial variability in downscaling results from the value perfect predictor experiment. *International Journal of Climatology*, 39(9), 3819–3845.
- Wilby, R.L., Conway, D. and Jones, P. (2002) Prospects for downscaling seasonal precipitation variability using conditioned weather generator parameters. *Hydrological Processes*, 16(6), 1215–1234.
- Wilby, R.L. and Wigley, T.M.L. (1997) Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography*, 21(4), 530–548.

- Wilks, D.S. and Wilby, R.L. (1999) The weather generation game: a review of stochastic weather models. *Progress in Physical Geography*, 23(3), 329–357.
- Zerenner, T., Venema, V., Friederichs, P. and Simmer, C. (2016) Downscaling near-surface atmospheric fields with multi-objective genetic programming. *Environmental Modeling and Software*, 84, 85–98.
- Zerenner, T., Venema, V., Friederichs, P. and Simmer, C. (2018) Deterministic and stochastic precipitation downscaling using multi-objective genetic programming. In: *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pp. New York, NY: ACM, 79–80.
- Zitzler, E. and Thiele, L. (1999) Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE Transactions on Evolutionary Computation*, 3(4), 257–271.
- Zorita, E., Hughes, J.P., Lettemaier, D.P. and von Storch, H. (1995) Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation. *Journal of Climate*, 8(5), 1023–1042.

How to cite this article: Zerenner, T., Venema, V., Friederichs, P., & Simmer, C. (2021). Multi-objective downscaling of precipitation time series by genetic programming. *International Journal of Climatology*, 41(14), 6162–6182. <https://doi.org/10.1002/joc.7172>