Statistical downscaling skill under present climate conditions

Maraun, Douglas; Widmann, Martin; Gutiérrez, José M.

DOI:
10.1002/joc.5877

Received: 6 April 2018 | Revised: 9 August 2018 | Accepted: 28 September 2018

DOI: 10.1002/joc.5877

International Journal of Climatology RMetS

SPECIAL ISSUE ARTICLE

Statistical downscaling skill under present climate conditions: A synthesis of the VALUE perfect predictor experiment

Douglas Maraun[1] | Martin Widmann[2] | José M. Gutiérrez[3]

[1]Wegener Center for Climate and Global Change, University of Graz, Graz, Austria

[2]School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, UK

[3]Meteorology Group, Instituto de Física de Cantabria, CSIC-University of Cantabria, Santander, Spain

Correspondence
Douglas Maraun, Head of the Regional Climate Research Group, Wegener Center for Climate and Global Change, University of Graz, Brandhofgasse 5, 8010 Graz, Austria.
Email: douglas.maraun@uni-graz.at

Funding information
Horizon 2020 Framework Programme, Grant/Award Number: COST Action ES1102

VALUE is a network that developed a framework to evaluate statistical downscaling methods including model output statistics such as simple bias correction and quantile mapping; perfect prognosis methods such as regression models and analog methods; and weather generators. The first experiment addresses the downscaling performance in present climate with perfect predictors. This paper presents a synthesis of the VALUE special issue, with a focus on the results of this first experiment. This paper presents a synthesis of the results. Model output statistics performs mostly well, but requires predictors at a resolution close to the target one. Perfect prog performance depends crucially on model structure and predictor choice. Weather generators perform in principle well for all aspects that can be expressed by the available model structure. Inter-annual variability is underrepresented by both perfect prog and weather generator approaches. Spatial variability is poorly represented by almost all participating methods (inherited by model output statistics from the driving model, not represented by the perfect prog and weather generator methods). Further studies are required to systematically assess (a) the role of predictor choice for perfect prog; (b) the performance of spatial weather generators, to study the performance based on GCM predictors; (c) downscaling skill in simulated future climates; and (d) the credibility of simulated predictors in a future climate.

KEYWORDS

bias correction, evaluation, regional climate, statistical downscaling, validation

1 | INTRODUCTION

Operational global coupled general circulation models (GCMs) such as those studied in the Coupled Model Intercomparison Project Phase 5 (CMIP5; Taylor *et al*., 2009) have a resolution too coarse to realistically represent the influence of regional-scale topography on climate, as well as regional-scale climatic phenomena themselves, in particular localized extreme events. This scale-gap is often bridged by downscaling (Giorgi and Mearns, 1999; Benestad, 2016; Maraun and Widmann, 2018), either by dynamical regional climate models (RCMs) or by statistical downscaling based on empirical relationships.

Regional climate projections are still a scientific challenge and require thorough evaluation (Nature, 2010; Barsugli *et al*., 2013; Hewitson *et al*., 2014; Maraun *et al*., 2015). The evaluation of downscaling-based regional climate projections requires (a) to assess how large-scale predictors are simulated and (b) to assess how well the downscaling itself performs. In a climate change context, both assessments have to address not only the performance in present climate, but also the performance to represent past climatic changes and potential future climates. The latter assessment of course can only be based on process-based plausibility arguments (Maraun and Widmann, 2018).

Regional climate phenomena have marginal (regarding the univariate unconditional distribution), temporal, spatial and inter-variable aspects (Maraun *et al*., 2015). Some aspects may specifically characterize extreme events, such as the tails of the marginal distribution or long spells. Most evaluation studies so far have only addressed marginal and to some extent temporal aspects, some with a focus on extreme events (Haylock *et al*., 2006; Goodess *et al*., 2010; Bürger *et al*., 2012). Spatial and inter-variable aspects have hardly been evaluated so far (Ferraris *et al*., 2003; Frost *et al*., 2011; Hu *et al*., 2013; Paschalis *et al*., 2013; Wilcke *et al*., 2013). Downscaling evaluation studies have typically focussed on a few methods, mostly from within one approach such as bias correction (Gudmundson *et al*., 2012; Teutschbein and Seibert, 2012; Gutmann *et al*., 2014). Until recently no comprehensive intercomparison and evaluation of different downscaling approaches existed.

The EU COST Action VALUE (Maraun *et al*., 2015) set out to systematically address this gap as far as possible. Three experiments have been designed: Experiment 1 to isolate downscaling skill in present climate by using observed (reanalysis-based, "perfect") predictors; Experiment 2 to assess the overall performance of GCM and downscaling in present climate; and Experiment 3 to assess downscaling skill in a future pseudo reality. In addition, the relevance of credible GCM projections has been assessed in a bias correction context (Maraun *et al*., 2017). As EU COST Action, VALUE received funding for travel and coordination only. The actual research was based on in-kind contributions. So far, VALUE has carried out Experiment 1. Owing to the limited capacity of the participants, inter-variable relationships have not been assessed yet. Similarly, it was not possible to systematically compare a range of different predictor selections for a specific method type. This special issue presents the VALUE results so far, including a discussion of the interface between climate modelling and users (Roessler *et al*., 2017), uncertainties resulting from observational data sets (Kotlarski *et al*., 2017; Herrera *et al*., submitted manuscript, 2018), and the evaluation results of the perfect predictor experiment (Gutiérrez *et al*., 2018; Maraun *et al*., 2018; Hertig *et al*., 2018; Soares *et al*., submitted manuscript, 2018; Widmann *et al*., submitted manuscript, 2018).

This short communication synthesizes the results across the special issue papers. The focus is on the results from the perfect predictor experiment, in particular regarding marginal, temporal, and spatial aspects, including extremal aspects. In addition to these aspects, also a process-oriented evaluation has been carried out. Given that the results for this assessment (Soares *et al*., submitted manuscript, 2018) are rather experimental and available only for selected climatic phenomena, we do not include them here in detail. Key messages regarding process-oriented evaluation, however, are discussed in the conclusions.

Details on Experiment 1 and the participating methods can be found in Gutiérrez *et al*. (2018). In particular, the predictors chosen for each method are listed therein. Note that Experiment 1 addresses the performance of downscaling methods in present climate only. A good performance in this experiment is not sufficient for a good performance in a future climate, let alone for an overall skilful regional climate projection.

2 | SYNTHESIS OF THE PERFECT PREDICTOR EXPERIMENT

Prior to discussing the results, we briefly review the different approaches of statistical downscaling. Depending on how statistical downscaling methods incorporate their predictors under calibration, they can be categorized into perfect prog (PP) and model output statistics (MOS; Rummukainen, 1997; Maraun *et al*., 2010; Maraun and Widmann, 2018).

A PP model is calibrated with both observed predictands and observed (here reanalysis-based) predictors. To generate regional future projections, the model is then applied to projections of the predictors as simulated by a climate model. Typical models are based on regression (including canonical correlation analysis [CCA] and nonlinear regression such as artificial neural networks), analogs and weather types. PP models have to fulfil three assumptions in a climate change context (Maraun and Widmann, 2018): (a) the PP-condition has to be fulfilled, that is, the predictors have to be realistically simulated in present climate, and credibly projected into the future. (b) Predictors need to be informative, that is, they have to explain a large fraction of local variability on all relevant time scales, including the response to climate change. (c) The structure of the statistical downscaling model has to be such that the influence of the predictors on the predictand is adequately represented for the aspects of interest, including at least moderate extrapolation to future climates.

In MOS, the model is calibrated between observed predictands, but simulated predictors. This approach intrinsically adjusts model biases. While the temporal synchrony between predictors and predictands in PP (because both are observed) allows for building regression models with a broad range of different predictor variables, MOS in a climate change context is typically much simpler: the climate model is not in synchrony with observations, such that only long-term distributions can be mapped. In a climate change context, MOS is thus typically restricted to bias correction of the simulated predictor variable. Widely used implementations are simple additive and scaling corrections or variants of more flexible quantile mapping. To be used for climate projections, MOS methods have to fullfil three assumptions (Maraun and Widmann, 2018): (a) the predictor needs to be realistically simulated in present climate, apart from correctable biases; under future conditions, the predictor has to be

credibly simulated. (b) The predictor needs to represent the predictand, that is, the same spatial scale and location. (3) The transfer function needs to have a suitable structure to represent the aspects of interest, and needs to be applicable in a future climate. The latter typically involves at least mild extrapolations to unobserved extremes.

Additionally weather generators (WGs) have been developed that can be implemented either as complex PP methods (when conditioned on predictors) or as so-called change factor weather generators. In general, WGs are stochastic models that explicitly model at least the marginal and temporal aspects of a meteorological variable, often even the relationships between a set of variables, sometimes also spatial dependence. In VALUE, all WGs are change-factor WGs, that is, they are used without predictors. Under future climate change, the parameters of such models would be adjusted according to changes simulated by a climate model. For these models, the assumptions are similar to those of MOS: (a) the change factors for the WG parameters have to be credibly simulated; (b) these simulated changes have to be representative of the changes at the WG location and scale; and (c) all relevant parameters that may change are modified by change factors.

The VALLUE perfect predictor experiment uses the ERA-Interim reanalysis (Dee *et al*., 2011) as driving data: either directly as input for the participating MOS methods and to calculate predictors for the PP methods; or as lateral boundary conditions for KNMI's RACMO2 RCM (van Meijgaard *et al*., 2008), which is used as alternative input for the MOS methods. The choice of reanalysis data as input ensures that boundary conditions and predictors are essentially bias free and—on inter-annual and longer timescales—synchronized with observations. This setup allows us to isolate the downscaling skill of the participating methods. As predictand data we choose 86 stations from the ECA-D data base (Klein Tank *et al*., 2002). The methods are calibrated and evaluated in a cross validation setup over the period 1979–2008. For details see Gutiérrez *et al*. (2018).

The VALUE perfect predictor experiment cannot address the PP and credibility assumptions: in this experiment, predictors are by construction perfectly simulated. It is designed to address the informativeness (PP)/representativeness (MOS) and the model structure assumptions under present climate conditions. A key issue, which is often overlooked for PP methods, is the fact that downscaling skill does not only depend on the chosen predictors, but also on the chosen model structure. In fact, the structure can often give information on model skill already prior to any evaluation (e.g., a deterministic regression model cannot capture extremes, which are not fully determined by the predictors; an analog method without additional adjustment of within-analog changes cannot represent thermodynamic changes (von Storch *et al*., 2000; Gutiérrez *et al*., 2013; Maraun and Widmann, 2018).

In the following we discuss the results of the perfect predictor experiment, separately for each downscaling approach and across the different aspects. Figures 1–3 summarize the performance of all participating methods for selected diagnostics for daily maximum (T_{max}) and minimum (T_{min}) temperature as well as precipitation. The diagnostics are briefly presented in Table 1. We do not discuss all methods individually, but focus on widely used implementations. For more in depth information, please refer to the individual papers of this special issue. We also do not discuss the representation of the seasonal cycle explicitly. It is calibrated for almost all methods, a good representation is thus trivial. Whether this calibration is still valid under future conditions is a separate issue which will be briefly discussed in the conclusions.

2.1 | Model output statistics

When driven with perfect and representative predictors, MOS methods should trivially improve all aspects they are calibrated for. The methods participating in VALUE all corrected marginal aspects only, some with a specific focus on extremes. Thus, almost all MOS methods perform well for essentially all marginal aspects for temperature, and the non-extreme aspects of precipitation. Regarding the tail of the precipitation distribution we obtained the following results: first, parametric models that explicitly model the tail outperform those based on a parametric distribution for the whole range of values (e.g., GPQM vs. GQM). Seasonally calibrated models perform better than those calibrated for the whole year (e.g., EQM vs. EQMs, but see the discussion in the conclusions). But finally, transfer functions with a constant extrapolation typically perform better than those with a parametric distribution (e.g., EQM vs. GPQM). The latter result to some extent questions the use of complex parametric extreme value models for the tail. But note that none of these approaches rests upon physical arguments. Whether any of these models are applicable under extrapolation to unobserved extremes under future climate conditions is essentially an open question.

Temporal variability is largely inherited by the driving model, it is only indirectly affected by modifications of the marginal distribution (e.g., by adjusting wet-day frequencies). The ERA and RCM performance for temperature is high regarding short term dependence and spells. For precipitation, the drizzle effect of ERA and the RCM is adjusted, resulting in a good representation of short-term dependence and spells. The representation of inter-annual variability by ERA-Interim and the RCM varies with season and variable. For temperature it is evident that the underestimation in spring by the RCM cannot be corrected by bias correction. For precipitation—in particular for more complex methods and almost independently of the performance of the driving data—bias correction often results in an overrepresentation of inter-annual variability. The effect is strongest in summer and is likely an inflation effect (Maraun, 2013).

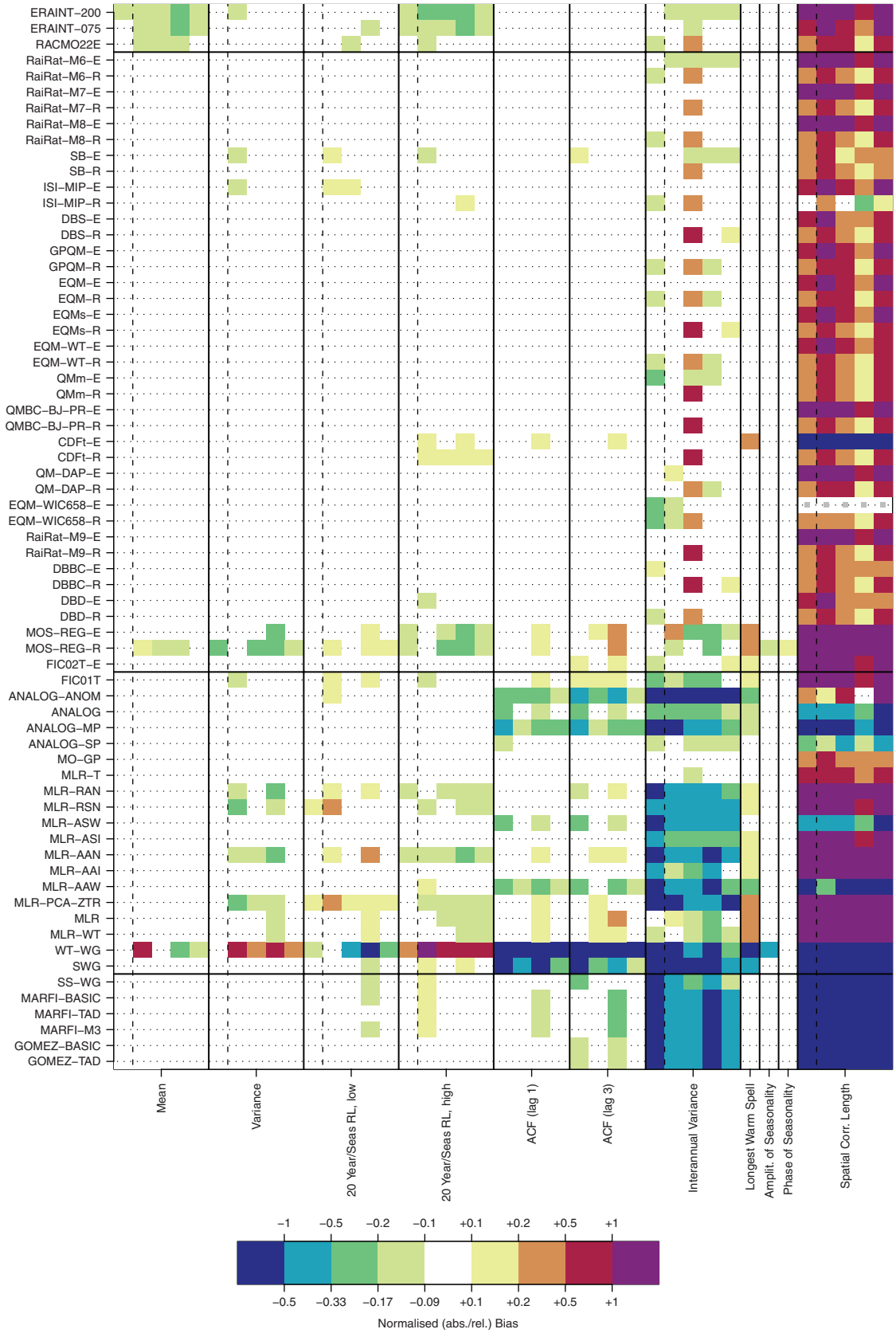

FIGURE 1 Results for daily maximum temperature T_{max}. Depending on the index, either the performance for the whole year (column separated by the dashed line) is shown, and/or all four seasons (four columns, from left to right: DJF, MAM, JJA, SON). In some cases the performance is evaluated only for the whole year. The definition of reference scales follows Maraun *et al.* (2018). Mean: twice the standard deviation of daily values; variance (daily and inter-annual), spell length, amplitude of seasonal cycle, spatial correlation length: the value itself; return level: absolute deviation from mean; correlations: 1; phase of the annual cycle: 1 month [Colour figure can be viewed at wileyonlinelibrary.com]

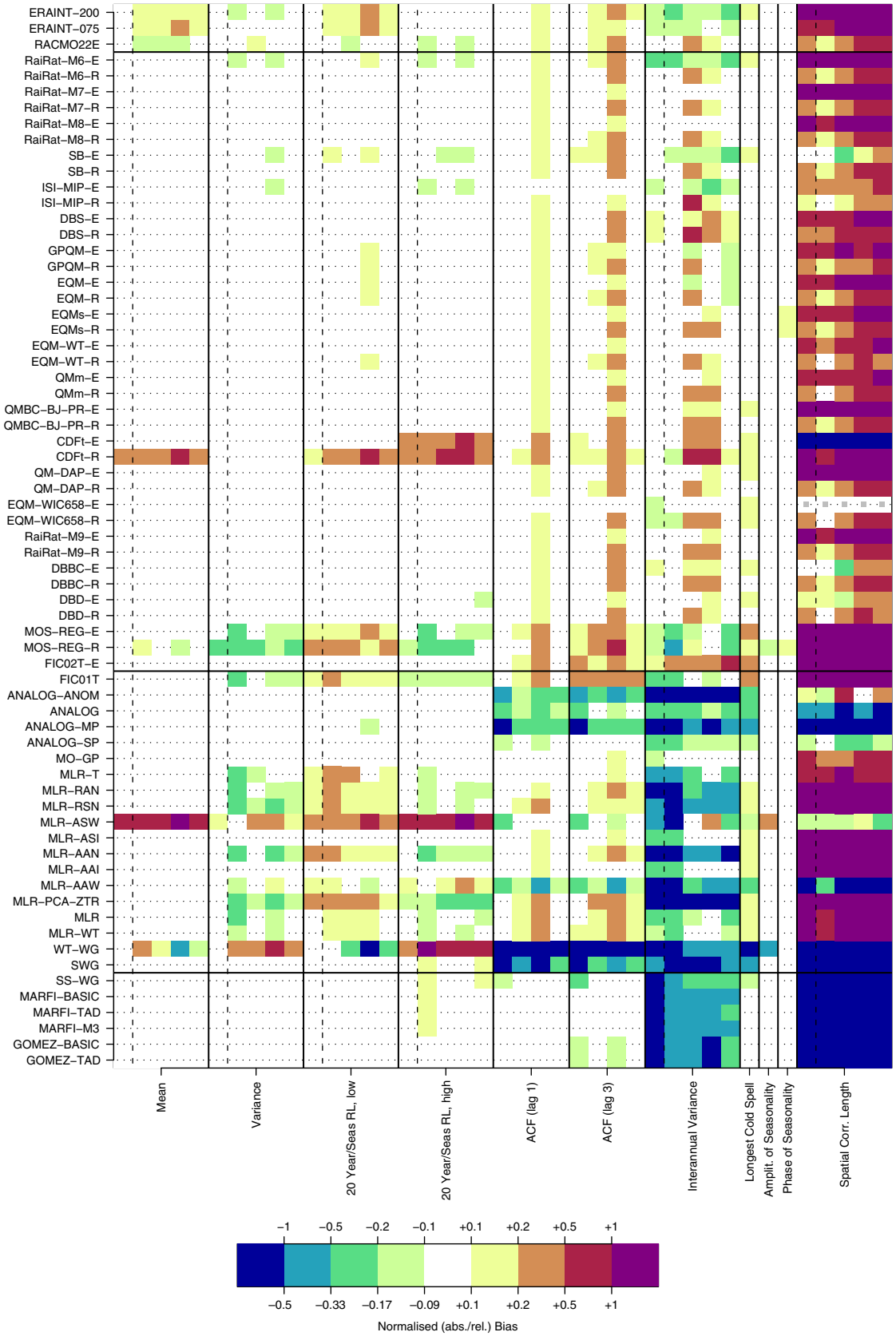

FIGURE 2 As Figure 1, but for daily minimum temperature T_{min}. The data uploaded to the VALUE portal for CDFt-R and MLR-ASW are most likely incorrect [Colour figure can be viewed at wileyonlinelibrary.com]

Also, spatial dependence is mostly inherited by the driving model. For example, the ERA-Interim temperature fields are far too smooth, an effect which is not corrected by univariate bias correction. Adjusting the marginal distribution has, however, an effect on precipitation fields by adjusting wet day frequencies. As a result, the added value of the

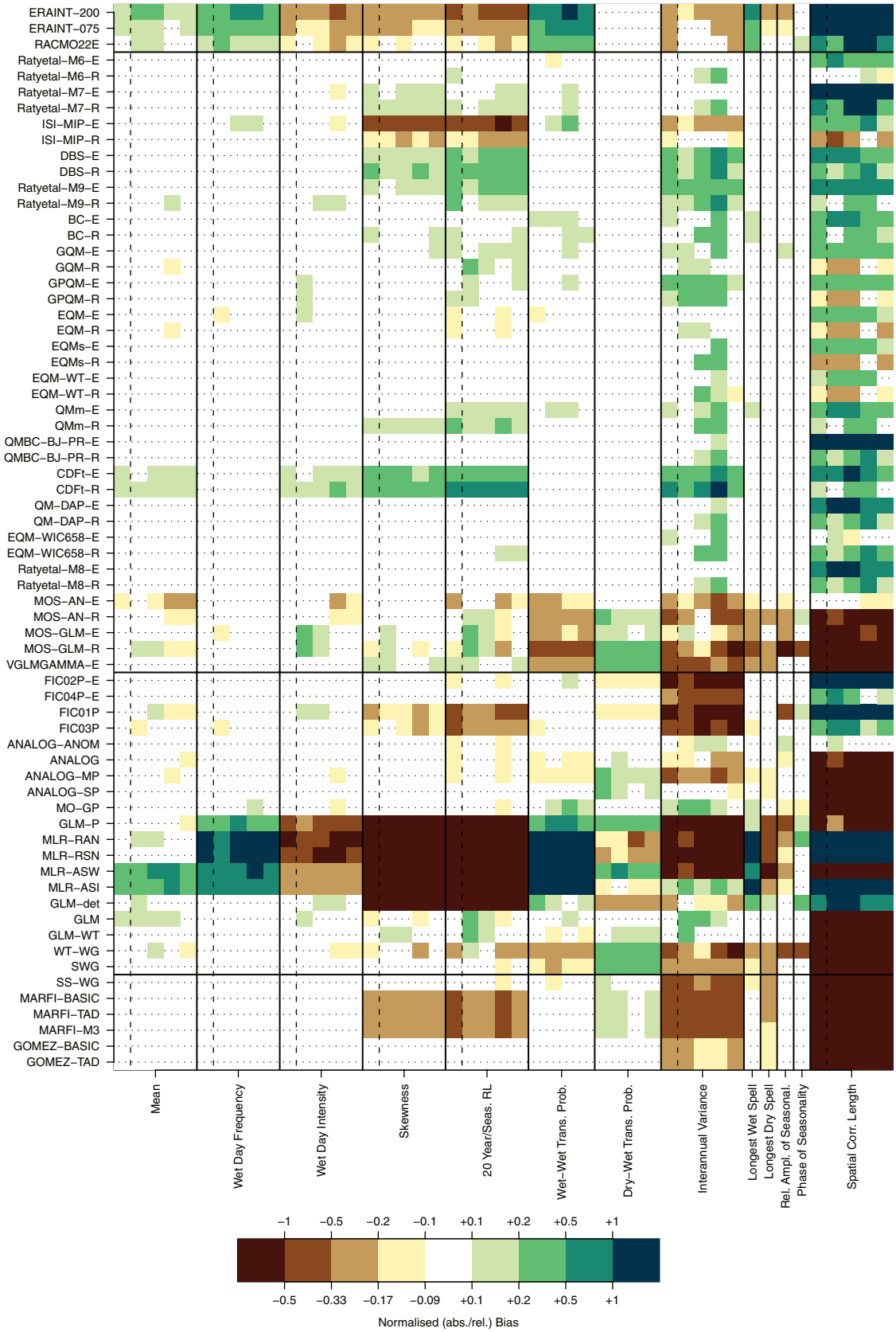

FIGURE 3 As Figure 1, but for precipitation. The definition of reference scales follows Maraun *et al*. (2018). For all indices apart from annual cycle phase: the value itself; phase of the annual cycle: 1 month [Colour figure can be viewed at wileyonlinelibrary.com]

RCM is crucial to improve spatial fields in particular for temperature, but in many cases also for precipitation. Bias correction methods that explicitly adjust the spatial–temporal structure would improve the evaluation results. But such a correction will by construction destroy the consistency with the driving model, the stronger, the worse the structure is

TABLE 1 Diagnostics considered in this synthesis. Further diagnostics are shown in the individual papers

Index	Variables	Performance measure	Resolution	Description
Marginal diagnostics				
Mean	T_{max}, T_{min} precipitation	Bias/rel. bias	Seasonal	Mean
Variance	T_{max}, T_{min} precipitation	Rel. bias	Seasonal	Variance
Skewness	Precipitation	Bias	Seasonal	Skewness
Wet day frequency	Precipitation	Bias	Seasonal	Number of wet days in year/season
Wet day intensity	Precipitation	Relative bias	Seasonal	Mean on wet days only
20-year/seas return level, low/high	T_{max}, T_{min} precipitation	Bias/rel. bias	Seasonal	20-year/season return level low/high: lower/upper tail of the distribution. Only upper tail for precipitation
Temporal diagnostics				
ACF1/3	T_{max}, T_{min}	Bias	Seasonal	Lag-1/lag-3 autocorrelation
Wet–wet transition prob.	Precipitation	Bias	Seasonal	Probability of a wet day, given that the previous day was wet
Dry–wet transition prob.	Precipitation	Bias	Seasonal	Probability of a dry day, given that the previous day was wet
Inter-annual variance	T_{max}, T_{min} precipitation	Rel. error	Seasonal	Variance of seasonally/annually averaged data
Longest warm spell	T_{max}	Bias	Seasonal	Median of the annual max. warm (>90th percentile) spell duration
Longest cold spell	T_{min}	Bias	Seasonal	Meadian of the annual max. cold (<10th percentile) spell duration
Longest wet spell	Precipitation	Bias	Seasonal	Median of the annual max. wet (≥1 mm) spell duration
Longest dry spell	Precipitation	Bias	Seasonal	Median of the annual max. dry (≤1 mm) spell duration
Amplitude of seasonality	T_{max}, T_{min}	Bias	Annual	Amplitude of the annual cycle
Rel. amplitude of season	Precipitation	Rel. error	Annual	Relative ampl. of the annual cycle
Phase of seasonality	T_{max}, T_{min}, precipitation	Bias	Annual	Day of the maximum of the smoothed seasonal cycle
Spatial diagnostics				
Spatial correlation length	T_{max}, T_{min}	Bias	Seasonal	Spatial distance at which the Pearson correlation between two series has decayed to 0.5
Spatial correlation length	Precipitation	Bias	Seasonal	Spatial distance at which the Spearman correlation between two series has decayed to 0.35

Note. For details see http://www.value-cost.eu/validationportal/app#!indices and click on "details" for the underlying R-code (note that registration is required).

simulated (Cannon, 2016; Maraun, 2016). This issue needs to be considered when applying such approaches to climate model output.

As discussed, bias correction largely inherits temporal and spatial variability from the driving model. Thus, the selection of appropriate driving models is crucial. Importantly, this is also a question of added value. Often it is argued that bias correction may be directly applied to a GCM as such an approach is cheaper than including an intermediate dynamical downscaling step. Our results show that this reasoning is in general wrong: of course, bias correction will trivially adjust present day climatologies to match observations at any available target scale (station, high-resolution grid). But RCMs resolve processes and small-scale variability below the GCM resolution that are crucial to represent short term persistence and spatial structure. Well performing RCMs may thus add crucial value as has been shown for, for example, short term persistence of daily maximum temperature (visible in Maraun *et al*., 2018) and spatial correlations of temperature (Widmann *et al*., submitted manuscript, 2018) (see also Figures 1 and 2). Of course, deficiencies of the chosen RCM may also deteriorate the performance of the driving model as has been the case, for example, for inter-annual variability of spring temperatures (Maraun *et al*., 2018; see also Figures 1 and 2). In short: bias correction cannot add value, but only climatological detail.

2.2 | Perfect prognosis

The performance of perfect prognosis depends strongly on the variable considered, the chosen method, and on the predictor choice. The mean of temperature and precipitation is essentially well represented for all methods.

Some implementations of the analog method have minor biases in mean temperature (see Gutiérrez *et al*., 2018, not visible in the summary plots) as the mean is not a calibrated statistic in the analog method. But in general the analog method represents marginal aspects well. Temporal dependence, however, is strongly underestimated for temperature, including long warm and cold spells and inter-annual variability. Here, algorithms sampling longer analogs or including a Markov component might help. For precipitation, the (anyhow weaker) dependence is well represented apart from

inter-annual variability. The representation of spatial dependence depends crucially on the implementation: if the analogs are defined simultaneously for several locations, dependence between these locations is well represented. If the analogs are defined only for single stations, spatial dependence is strongly underrepresented. Here a trade-off is necessary between tailoring predictors (which define the analog) for a small area, and representing spatial dependence over a large area. Classical analog methods, however, are strongly limited in representing long-term forced climatic changes (Gutiérrez *et al*., 2013). They sample from observed analogs and cannot represent climatic states with strongly altered thermodynamic conditions. For example, a certain circulation type will have a typical temperature in present climate, but a much warmer temperature in a future climate (and similarly more intense precipitation). Some authors suggest "constructed" analogs to create such unobserved future analogs (Maurer and Hidalgo, 2008).

Simple linear regression without randomisation modestly misrepresents marginal aspects (apart from the calibrated mean) of temperature, indicating that much of the local variability is explained by the predictors. Here, the dependence depends considerably on the chosen predictors: some implementations (MLT-T) use grid-box surface temperature as predictor and thereby obtain favourable results. These predictors, however, will likely not fullfill the perfect prog condition, that is, the performance would strongly drop with GCM simulated predictors. Inflation apparently improves temperature variance and high quantiles. But this approach is ill-designed and wrongly assumes that all local variance is explained by large-scale predictors. The inflation problem is mostly visible in the temporal correlation: it is by construction not improved compared to the deterministic regression as no local variability is simulated.[1] Linear regression or any kind of deterministic regression (including inflation) fails to downscale daily precipitation: apart from the mean essentially all aspects are badly represented. The main reason is that the predictors explain a rather low fraction of local precipitation variability and do not represent the skewed process. A generalized linear model with randomisation (GLM), however, performs well for almost all aspects. In particular, also temporal aspects are well represented even though no dependence conditional on the predictors is modelled. This finding indicates that most of the precipitation dependence does not result from any direct dependence between subsequent precipitation events, but is rather imprinted by the large-scale circulation. All regression models participating in VALUE have difficulties representing spatial dependence: the deterministic implementations are too smooth in space (in particular for precipitation) because they do not simulate local variability; the stochastic methods with randomisation do not simulate the dependence between stations and are thus too noisy in space. Here, an explicit spatial dependence model would be necessary.

In contrast to simple bias correction methods (the situation is different for regression-based MOS in weather forecasting), PP includes information on physical processes and thus can in principle added value to the driving model. For instance, the analog method models spatial dependence conditionally on specific large-scale weather types.

2.3 | Weather generators

In principle, weather generators can represent any aspect they are calibrated for. This general statement is of course strongly limited by the availability of a sufficiently complex statistical model, and by the amount of data to constrain the model structure and parameters. In general, however, marginal aspects are well represented. For temperature, a Gaussian distribution seems to be sufficient, although high and low extremes are not very well captured in summer and winter, respectively. For precipitation, a double exponential distribution (SS-WG) seems to suffice, a simple gamma distribution does not fully represent skewness and extremes (the MARFI WGs). Nonparametric resampling-based models, of course, perform well for marginal aspects (GOMEZ). Temperature and precipitation spells are well represented apart from long dry spells. Inter-annual variability is, as often noted in the literature, strongly underrepresented. Here, conditioning on atmospheric predictors may improve the results (Katz and Parlange, 1993; Wilks and Wilby, 1999), although the corresponding findings for perfect prog demonstrate the limitations of this idea. None of the weather generators employed an explicit spatial model and therefore fully miss to represent spatial dependence. More advanced models are required such as multisite Richardson-type or truncated Gaussian weather generators (Bárdossy and Plate, 1992; Wilks, 1998; Ferraris *et al*., 2003; Paschalis *et al*., 2013), spatial GLMs (Yang *et al*., 2005), non-homogeneous hidden Markov models (Hughes and Guttorp, 1994), spatial Poisson cluster models (Cox and Isham, 1994; Northrop, 1998) or random cascade weather generators (Schertzer and Lovejoy, 1987; Thober *et al*., 2014). Knowledge of their actual performance in practical applications is still limited and often contradictory (Frost *et al*., 2011; Hu *et al*., 2013). Whether weather generators can add value or not depends on their setup: in a PP setting they can in principle (see discussion in section 2.2), in a change factor setting they do not include any process information and thus—as MOS (section 2.1)—can only add climatological detail.

3 | DISCUSSION AND CONCLUSIONS

We have presented a synthesis of the VALUE perfect predictor evaluation experiment for statistical downscaling methods. The main results are:

- With perfect predictors, MOS is performing best in present climate, in particular flexible quantile mapping approaches. MOS can rather easily be applied over large areas, but relies on skilfully simulated predictors with a resolution close to the target resolution. In other words: standard MOS cannot by itself bridge a scale-gap, and dynamical downscaling could potentially add substantial value. This is most clear for spatial dependence, where the bias correction of the RCM is much more skilful than the bias correction of the reanalysis. Thus, MOS applied to GCM data will produce information representing the GCM resolution, and also MOS applied to RCM data will not represent point scale values.
- The performance of perfect prognosis is generally lower and depends strongly on the method type and the chosen predictors. Here, a good understanding and choice of model structure for a given purpose is crucial. Essentially all method types fail to reproduce inter-annual and spatial variability. A key issue that has to be considered is the perfect prog condition: grid-box surface predictors may be well simulated in the reanalysis and will result in high predictive power. But they will likely be poorly represented by a free running GCM, resulting in low downscaling skill. If a realistic representation of spatial–temporal variability is required, perfect prog has its strength in producing tailored output for a small domain. For large areas, a tailored predictor selection will be costly, and spatial models will be computationally expensive and likely infeasible; here, perfect prog may provide a cheap way to generate local information of mean climate from large ensembles.
- Weather generators with the right model for the marginal distribution perform well for all aspects apart from inter-annual variability, long dry spells and spatial variability. For the latter aspect, further research is required to assess the skill of specifically designed multi-site weather generators. The strength of weather generators is in producing either single-site information (also over many sites) with characteristics very close to the observed, or—potentially—to produce tailored spatial–temporal models over a small domain.

Seasonality in present climate is well captured by most methods—it is explicitly modelled either by monthly or seasonal training, or by adding a parametric cycle. Such models trivially perform well in present conditions, but may not capture changes in the seasonal cycle. In fact, seasonally varying biases indicate that biases may also change in a future climate. Some methods therefore attempt to describe the seasonal cycle by atmospheric predictors. Such predictors have to account not only for variations of circulation, but also for thermodynamic variations (e.g., changing moisture content) throughout the year.

A key question in downscaling is that of added value. Bias correction does not incorporate any process information and thus can only add climatological detail[2]. As a consequence, the performance of a bias corrected climate model simulations depends strongly on the chosen climate model. In particular for representing temporal and spatial variability, VALUE has demonstrated that dynamical downscaling has the potential to crucially add value: for instance, a GCM in standard resolution will not realistically represent the spatial dependence structure of precipitation events. An RCM may prove to be much more realistic. Soares *et al*. (submitted manuscript, 2018) found similar results in the process-oriented evaluation: if the sensitivity of local weather to relevant weather phenomena (such as the NAO, synoptic weather patterns, or regional phenomena such as Foehn winds) is not represented by the driving model, bias correction cannot generate this sensitivity. Perfect prognosis incorporates process information and may in principle add value. For instance, the analog method links spatial dependence to large-scale weather types and thus improves the corresponding GCM representation. However, Soares *et al*. (submitted manuscript, 2018) found that in practice, perfect prognosis methods often do not realistically incorporate the sensitivity of local weather to regional-scale weather phenomena. In terms of added value, change factor weather generators behave as bias correction: it can only add climatological detail.

In addition to the evaluation of downscaling methods, VALUE also addressed the quality of observational datasets used as reference for the evaluation (Kotlarski *et al*., 2017; Herrera *et al*., submitted manuscript, 2018). Kotlarski *et al*. (2017) evaluated a suite of RCMs against three different gridded reference data sets. They found that the uncertainty inherent in these datasets was typically smaller than climate model uncertainty. For individual regions and seasons, however, the ranking of the different climate models depended on the choice of the reference data set. Herrera *et al*. (submitted manuscript, 2018) analysed the influence of station density, interpolation method and spatial resolution on gridded data sets. They found that a sufficient station density of about six stations per grid box is crucial to obtain a good representation of area average statistics. These results highlight the relevance of high-quality reference data for climate model evaluation at the regional scale.

In the VALUE perfect predictor experiment we have not systematically investigated the influence of different perfect prog predictors on the downscaling skill (although some conclusions could be drawn in Maraun *et al*. (2018). Furthermore, no models have participated that explicitly simulate spatial dependence. Both aspects are important though and require further research.

The results from the considered VALUE experiment only hold for perfect predictors in present climate. So far, we have not assessed (a) how well the predictors are simulated

by climate models, (b) how well the downscaling methods perform under future climate conditions, and (c) how credible the driving climate models simulate the predictors in a future climate. Issues (a) and (b) are planned to be assessed in further experiments (in the framework of EURO-CORDEX, where VALUE activities have been merged). For the different downscaling approaches, these issues specifically involve the following questions raised in Maraun and Widmann (2018):

- For MOS: which model biases are correctable at all? Is the model output representative of the observations and simulates local forcings and feedbacks? Is the model structure, in particular of different quantile mapping variants, suitable to account for changes in model biases? To what extent are multivariate MOS approaches feasible and defensible?
- For PP: is the perfect prog condition fulfilled, that is, are the predictors realistically simulated in present climate, and credibly projected into the future? Are all predictors necessary to describe climatic changes in the aspects of interest included? Is the model structure appropriate to describe the interplay of predictors, and their changes? The issue of predictor selection and model building is highly non-trivial as models are calibrated on short-term variability, but applied to long-term variability, such that standard statistical procedures are strictly speaking not valid.
- For change factor weather generators: are all relevant parameters describing marginal, temporal and spatial aspects that might change in a different climate modified by change factors? Are the simulated change factors representative of the location (as in MOS, the simulated area average climate change signal might not be representative of the climate change signal at the target scale)?

A key issue is that the evaluation of regional climate simulations has to address GCM performance, in particular regarding the credibility of future projections of the predictors (either directly in the GCM, or after dynamical downscaling). Such an assessment should rest upon two pillars: first it should be assessed whether observed and simulated predictor trends in present climate are consistent, that is, indistinguishable apart from internal variability (Bhend and Whetton, 2013); and second, it should be assessed whether the processes controlling changes in the predictors are realistically simulated (Maraun *et al*., 2017).

All these issues are relevant for the construction of regional ensemble projections based on statistical downscaling: first, GCMs should be selected that simulate realistic and credible predictors for PP, representative and credible predictors for MOS, or representative and credible change factors for weather generators. In case the predictors or change factors are not representative, one should consider dynamical downscaling. Second, one should only select well performing statistical downscaling methods that are designed for the purpose of interest—both in terms of including all relevant predictors or change factors, and in terms of having an appropriate model structure.

Regional climate change projections are often developed and provided in the context climate change impact modelling and decision making. Surveys such as that conducted within VALUE (Roessler *et al*., 2017) revealed that users often have rather unrealistic data demands far exceeding what is currently feasible and defensible. Here a close communication, knowledge exchange and negotiations of what can and what cannot be provided: climate modellers have to understand user needs, and provide sensible—possibly alternative—options. Users have to understand model limitations and develop approaches that can cope with these limitations. Roessler *et al*. (2017) discuss knowledge gaps, communication gaps, and structural gaps hampering this process. As key issues they call for sensible guidelines describing model limitations and uncertainties in an honest and transparent way; face-to-face communication between modellers and users; and mechanisms to finance the collaboration between modellers and users throughout the process of knowledge production.

ACKNOWLEDGEMENTS

VALUE was funded from 2012 to 2015 as EU COST Action ES1102.

NOTES

[1]In addition, the inflation problem should be visible in the interannual variability, but here the underestimation of the regression method dominates the inflation effect.

[2]Note that this is different in weather forecasting: here, MOS includes process information via meteorological predictors and may well add value

ORCID

Douglas Maraun http://orcid.org/0000-0002-4076-0456
Martin Widmann http://orcid.org/0000-0001-5447-5763
José M. Gutiérrez http://orcid.org/0000-0002-2766-6297

REFERENCES

Bárdossy, A. and Plate, E.J. (1992) Space-time model for daily rainfall using atmospheric circulation patterns. *Water Resources Research*, 28, 1247–1259.

Barsugli, J.J., Guentchev, G., Horton, R.M., Wood, A., Mearns, L.O., Liang, X.-Z., Winkler, J.A., Dixon, K., Hayhoe, K., Rood, R.B., Goddard, L., Ray, A., Buja, L. and Ammann, C. (2013) The practitioner's dilemma: how to assess the credibility of downscaled climate projections. *Eos, Transcations American Geophysical Union*, 94(46), 424–425.

Benestad, R. (2016) *Downscaling Climate Information*. Oxford: Oxford Research Encyclopedia of Climate Science. https://doi.org/10.1093/acrefore/9780190228620.013.27.

Bhend, J. and Whetton, P. (2013) Consistency of simulated and observed regional changes in temperature, sea level pressure and precipitation. *Climate Change*, 118(3–4), 799–810.

Bürger, G., Murdock, T.Q., Werner, A.T., Sobie, S.R. and Cannon, A.J. (2012) Downscaling extremes—an intercomparison of multiple statistical methods for present climate. *Journal of Climate*, 25, 4366–4388.

Cannon, A.J. (2016) Multivariate bias correction of climate model output: matching marginal distributions and intervariable dependence structure. *Journal of Climate*, 29(19), 7045–7064.

Cox, D.R. and Isham, V.S. (1994) Stochastic models of precipitation. In: *Statistics for the Environment, 2: Water Related Issues*. Chichester and New York, NY: John and Wiley, pp. 3–18.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beeljars, A.C.M., van den 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, L., 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, 553–597.

Ferraris, L., Gabellani, S., Rebora, N. and Provenzale, A. (2003) A comparison of stochastic models for spatial rainfall downscaling. *Water Resources Research*, 39(12), 1368.

Frost, A.J., Charles, S.P., Timbal, B., Chiew, F.H.S., Mehrotra, R., Nguyen, K. C., Chandler, R.E., McGregor, J.L., Fu, G., Kirono, D.G.C., Fernandez, E. and Kent, D.M. (2011) A comparison of multi-site daily rainfall downscaling techniques under Australian conditions. *Journal of Hydrology*, 408(1), 1–18.

Giorgi, F. and Mearns, L.O. (1999) Introduction to special section: regional climate modeling revisited. *Journal of Geophysical Research*, 104(D6), 6335–6352.

Goodess, C.M., Anagnostopoulou, C., Bárdossy, A., Frei, C., Harpham, C., Haylock, M.R., Hundecha, Y., Maheras, P., Ribalaygua, J., Schmidli, J., Schmith, T., Tolika, K., Tomozeiu, R. and Wilby, R.L. (2010) *An intercomparison of statistical downscaling methods for Europe and European regions—assessing their performance with respect to extreme weather events and the implications for climate change applications*. Norwich: Climatic Research Unit. Technical report.

Gudmundson, L., Bremnes, J.B., Haugen, J.E. and Engen-Skaugen, T. (2012) Technical note: downscaling RCM precipitation to the station scale using statistical transformations—a comparison of methods. *Hydrology and Earth System Sciences*, 16, 3383–3390.

Gutiérrez, J.M., San-Martín, D., Brands, S., Manzanas, R. and Herrera, S. (2013) Reassessing statistical downscaling techniques for their robust application under climate change conditions. *Journal of Climate*, 26(1), 171–188.

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., Manzanas, R., Iturbide, M., Vrac, M., Dubrovsky, M., Ribalaygua, J., Pórtoles, J., Räty, 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. (2018) An intercomparison of a large ensemble of statistical downscaling methods for Europe: overall results from the VALUE perfect predictor cross-validation experiment. *International Journal of Climatology*. https://doi.org/10.1002/joc.5462.

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.

Haylock, M.R., Gawley, G.C., Harpham, C., Wilby, R.L. and Goodess, C.M. (2006) Downscaling heavy precipitation over the United Kingdom: a comparison of dynamical and statistical methods and their future scenarios. *International Journal of Climatology*, 26(10), 1397–1415.

Hertig, E., Maraun, D., Bartholy, J., Pongracz, R., Vrac, M., Mares, I., Gutiérrez, J.M., Wibig, J., Casanueva, A., Soares, P.M.M. (2018) Validation of extremes from the Perfect-Predictor Experiment of the COST Action VALUE. *International Journal of Climatology*. https://doi.org/10.1002/joc.5469.

Herrera, S., Kotlarski, S., Soares, P.M.M., Cardoso, R.M., Jaczewski, A., Gutiérrez, J.M. and Maraun, D. (2018) Uncertainty in gridded precipitation products: Influence of station density, interpolation method and grid resolution. *International Journal of Climatology*, https://doi.org/10.1002/joc.5878.

Hewitson, B.C., Daron, J., Crane, R.G., Zermoglio, M.F. and Jack, C. (2014) Interrogating empirical–statistical downscaling. *Climate Change*, 122, 539–554.

Hu, Y., Maskey, S. and Uhlenbrook, S. (2013) Downscaling daily precipitation over the Yellow River source region in China: a comparison of three statistical downscaling methods. *Theoretical and Applied Climatology*, 112(3–4), 447–460.

Hughes, J.P. and Guttorp, P. (1994) A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resources Research*, 30(5), 1535–1546.

Katz, R.W. and Parlange, M.B. (1993) Effects of an index of atmospheric circulation on stochastic properties of precipitation. *Water Resources Research*, 29(7), 2335–2344.

Klein Tank, A.M.G., 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., Mietus, M., Coelho, F., Mares, C., Razuvaev, V., Nieplova, E., Cegnar, T., López, J.A., Dahlström, B., Moberg, A., Kirchhofer, W., Ceylan, A., Pachaliuk, O., Alexander, L.V. and Petrovic, P. (2002) Daily dataset of 20th-century surface air temperature and precipitation series for the European climate assessment. *International Journal of Climatology*, 22(12), 1441–1453.

Kotlarski, S., Szabó, P., Herrera, S., Räty, O., Keuler, K., Soares, P.M.M., Cardoso, R.M., Bosshard, T., Pagé, C., Boberg, F., Gutiérrez, J.M., Isotta, F. A., Jaczewski, A., Kreienkamp, F., Liniger, M.A., Lussana, C. and Pianko-Kluczynska, K. (2017) Observational uncertainty and regional climate model evaluation: a pan-European perspective. *International Journal of Climatology*. https://doi.org/10.1002/joc.5249.

Maraun, D. (2013) Bias correction, quantile mapping and downscaling: revisiting the inflation issue. *Journal of Climate*, 26, 2137–2143.

Maraun, D. (2016) Bias correcting climate change simulations—a critical review. *Current Climate Change Reports*, 2(4), 211–220. https://doi.org/10.1007/s40641-016-0050-x.

Maraun, D. and Widmann, M. (2018) *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge: Cambridge University Press.

Maraun, D., Wetterhall, F., Ireson, A.M., Chandler, R.E., Kendon, E.J., Widmann, M., Brienen, S., Rust, H.W., Sauter, T., Themeßl, M., Venema, V.K.C., Chun, K.P., Goodess, C.M., Jones, R.G., Onof, C., Vrac, M. 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, RG3003.

Maraun, D., Widmann, M., Gutierrez, J.M., Kotlarski, S., Chandler, R.E., Hertig, E., Wibig, J., Huth, R. and Wilcke, R.A.I. (2015) VALUE: a framework to validate downscaling approaches for climate change studies. *Earth's Future*, 3, 1–14.

Maraun, D., Shepherd, T.G., Widmann, M., Zappa, G., Walton, D., Hall, A., Gutierrez, J.M., Hagemann, S., Richter, I., Soares, P. and Mearns, L. (2017) Towards process-informed bias correction of climate change simulations. *Nature Climate Change*, 7(11), 764.

Maraun, D., Huth, R., Gutierrez, J.M., San Martin, D., Dubrovsky, M., Fischer, A., Hertig, E., Soares, P.M., Bartholy, J., Pongracz, R., Widmann, M., Casado, M.J. and Ramos, P. (2018) The VALUE perfect predictor experiment: evaluation of temporal variability. *International Journal of Climatology*. https://doi.org/10.1002/joc.5222.

Maurer, E.P. and Hidalgo, H.G. (2008) *Utility of daily vs. monthly large-scale climate data: an intercomparison of two statistical downscaling methods*. Santa Clara: Santa Clara University. Technical report.

van Meijgaard, E., van Ulft, L.H., van de Berg, W.J., Bosveld, F.C., van den Hurk, B.J.J.M., Lenderink, G. and Siebesma, A.P. (2008) *The KNMI regional atmospheric climate model RACMO version 2.1*. De Bilt: Royal Dutch Meteorological Institute (KNMI). Technical report number: 302.

Nature. (2010) Validation required. *Nature*, 463(7283), 849–849. https://doi.org/10.1038/463849a.

Northrop, P. (1998) A clustered spatial-temporal model of rainfall. *Proceedings of the Royal Society A*, 454(1975), 1875–1888.

Paschalis, A., Molnar, P., Fatichi, S. and Burlando, B. (2013) A stochastic model for high-resolution space-time precipitation simulation. *Water Resources Research*, 49(12), 8400–8417.

Roessler, O., Fischer, A.M., Huebener, H., Maraun, D., Benestad, R.E., Christodoulides, P., Soares, P.M.M., Cardoso, R.M., Pagé, C., Kanamaru, H., Kreienkamp, F. and Vlachogiannis, D. (2017) Challenges to link climate change data provision and user needs—perspective from the COST-Action VALUE. *International Journal of Climatology*. https://doi.org/10.1002/joc.5060.

Rummukainen, M. (1997) *Methods of statistical downscaling of GCM simulations. Reports meteorology and climatology*. Norrköping: Swedish Meteorological and Hydrological Institute. Technical report 80, SE-601 76.

Schertzer, D. and Lovejoy, S. (1987) Physical modeling and analysis of rain and clouds by anisotropic scaling multiplicative processes. *Journal of Geophysical Research*, 92(D8), 9693–9714.

von Storch, H., Langenberg, H. and Feser, F. (2000) A spectral nudging technique for dynamical downscaling purposes. *Monthly Weather Review*, 128, 3664–3673.

Taylor, K.E., Stouffer, R.J. and Meehl, G.A. (2009) A summary of the CMIP5 experiment design. Available at: http://cmip-pcmdi.llnl.gov/-cmip5/docs/Taylor_CMIP5_design.pdf.

Teutschbein, C. and Seibert, J. (2012) 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.

Thober, S., Mai, J., Zink, M. and Samaniego, L. (2014) Stochastic temporal disaggregation of monthly precipitation for regional gridded data sets. *Water Resources Research*, 50(11), 8714–8735.

Wilcke, R.A.I., Mendlik, T. and Gobiet, A. (2013) Multi-variable error correction of regional climate models. *Climate Change*, 120(4), 871–887.

Wilks, D.S. (1998) Multisite generalization of a daily precipitation generation model. *Journal of Hydrology*, 210, 178–191.

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.

Yang, C., Chandler, R.E. and Isham, V.S. (2005) Spatial-temporal rainfall simulation using generalized linear models. *Water Resources Research*, 41, W11415.

How to cite this article: Maraun D, Widmann M, Gutiérrez JM. Statistical downscaling skill under present climate conditions: A synthesis of the VALUE perfect predictor experiment. *Int J Climatol*. 2018; 1–12. https://doi.org/10.1002/joc.5877