AGU PUBLICATIONS

Earth's Future

RESEARCH ARTICLE

10.1002/2014EF000259

VALUE: A framework to validate downscaling approaches for climate change studies

Douglas Maraun[1], **Martin Widmann**[2], **José M. Gutiérrez**[3], **Sven Kotlarski**[4], **Richard E. Chandler**[5], **Elke Hertig**[6], **Joanna Wibig**[7], **Radan Huth**[8], **and Renate A.I. Wilcke**[9]

[1]GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany, [2]School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, UK, [3]Institute of Physics of Cantabria, IFCA, Santander, Spain, [4]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland, [5]Department of Statistical Science, University College London, London, UK, [6]Institute of Geography, University of Augsburg, Augsburg, Germany, [7]Department of Meteorology and Climatology, University of Lodz, Lodz, Poland, [8]Department of Physical Geography and Geoecology, Faculty of Science, Charles University and Institute of Atmospheric Physics, Academy of Sciences of the Czech Republic, Prague, Czech Republic, [9]Rossby Centre, Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

Key Points:

- VALUE has developed a framework to validate and compare downscaling methods
- The experiments comprise different observed and pseudo-reality reference data
- The framework is the basis for a comprehensive downscaling comparison study

Corresponding author:
Douglas Maraun, dmaraun@geomar.de

Citation:
Maraun, D., M. Widmann, J. M. Gutiérrez, S. Kotlarski, R. E. Chandler, E. Hertig, J. Wibig, R. Huth, and R. A. I. Wilcke (2015), VALUE: A framework to validate downscaling approaches for climate change studies, *Earth's Future*, *3*, 1–14, doi:10.1002/2014EF000259.

Received 10 JUN 2014
Accepted 26 NOV 2014
Accepted article online 4 DEC 2014
Published online 7 JAN 2015

Abstract VALUE is an open European network to validate and compare downscaling methods for climate change research. VALUE aims to foster collaboration and knowledge exchange between climatologists, impact modellers, statisticians, and stakeholders to establish an interdisciplinary downscaling community. A key deliverable of VALUE is the development of a systematic validation framework to enable the assessment and comparison of both dynamical and statistical downscaling methods. In this paper, we present the key ingredients of this framework. VALUE's main approach to validation is user- focused: starting from a specific user problem, a validation tree guides the selection of relevant validation indices and performance measures. Several experiments have been designed to isolate specific points in the downscaling procedure where problems may occur: what is the isolated downscaling skill? How do statistical and dynamical methods compare? How do methods perform at different spatial scales? Do methods fail in representing regional climate change? How is the overall representation of regional climate, including errors inherited from global climate models? The framework will be the basis for a comprehensive community-open downscaling intercomparison study, but is intended also to provide general guidance for other validation studies.

1. Introduction

The need for adaptation to the impacts of a changing climate has long been recognized. The United Nations Framwork Convention on Climate Change from 1992 states that "all parties shall cooperate in preparing for adaptation to the impacts of climate change; develop and elaborate appropriate and integrated plans for coastal zone management, water resources and agriculture, and for the protection and rehabilitation of areas […]" [*United Nations*, 1992]. In response, the World Meteorological Organisation established the Global Framework for Climate Services in 2009 (www.gfcs-climate.org), and several countries have developed adaptation strategies. For instance, the US Environmental Protection Agency released a draft climate change adaptation plan for public comment in February 2013 (www.epa.gov/climatechange/impacts-adaptation/ fed-programs.html), and the European Commission adopted the EU strategy on adaptation to climate change in April 2013 [*European Commission*, 2013]. The EU strategy acknowledges the need to take adaptation measures at all levels ranging from national to regional and local levels. Yet at the same time, it highlights the need for bridging key knowledge gaps, in particular regional and local-level analyses and risk assessments.

Scenarios of regional climate change are in general based on downscaled atmosphere ocean coupled general circulation models (AOGCMs, in the following simply referred to as GCMs), also called global climate models. The downscaling is either based on regional climate models [RCMs, *Rummukainen*, 2010], statistical methods [*Fowler et al.*, 2007; *Maraun et al.*, 2010], or a combination of both [*Maraun et al.*, 2010].

Barsugli et al. [2013] and *Hewitson et al.* [2013] point out that users of regional climate change scenarios are not hindered by a lack of, but rather, the abundance of available downscaling studies, often with highly uncertain or even contradictory results. *Hewitson et al.* [2013] emphasize the ethical dimension of providing climate change projections that are used in impact and adaptation projects. If taken seriously, climate projections will be used as the basis for decisions to spend real money, to make real changes to infrastructure, society and the environment, that affect real people and ecosystems.

In the light of our limited knowledge of regional climate change, debates have flared up whether state-of-the-art climate models are actually ready to provide input for impact studies [*Kundzewicz and Stakhiv*, 2010; *Wilby*, 2010] and how to include information from (regional) climate simulations into adaptation planning [*Lempert et al.*, 2004; *Prudhomme et al.*, 2010; *Wilby and Dessai*, 2010; *Brown et al.*, 2012]. Approaches have been advocated that emphasize the importance of first screening different adaptation options for different types of vulnerabilities; only if options turn out to be vulnerable to climate change, downscaling studies are potentially employed to identify robust options [e.g., *Lempert et al.*, 2004; *Wilby and Dessai*, 2010; *Brown et al.*, 2012]. *Dessai* [2009] and *Wilby and Dessai* [2010] argue that many options can be implemented without detailed knowledge of climate change. The scientific questions of skill are of course related to questions regarding the decision process, but should be kept mentally separate. In many cases, regional climate change information is relevant for adaptation planning. Also in many cases, downscaling does add value to GCM simulations in representing regional climate [e.g., *Deser et al.*, 2011]. In specific cases such as extreme summer precipitation, advanced downscaling methods might even be essential to correctly represent key processes and therefore the correct climate change [e.g., *Kendon et al.*, 2014].

Our current knowledge—despite a wealth of experience of international bodies such as the world bank [*Worldbank Independent Evaluation Group*, 2012]—is limited about where and to what extent state-of-the-art regional climate change scenarios add value to adaptation planning. In general, the answer depends on the region of interest, the variable of interest, the time horizon of interest, and the skill required by the particular user. Decisions that consider regional projections of change, accounting comprehensively and appropriately for the uncertainties in these projections, will be more robust and cost effective in many cases , as well as procedurally more defensible, than decisions that neglect this information. The precise question therefore is where, for which variables, and for which particular adaptation problem downscaling provides actionable and defensible information.

To guide practitioners with credible regional climate change simulations, a thorough understanding of their uncertainties and limitations is therefore indispensible. Key limitations of individual products, their applicability in different contexts, as well as differences between products should be made publicly available in an easily accessible way. Such publications need to go beyond previous qualitative and unspecific—and thereby potentially misleading—inventories [e.g., *UNFCCC*, 2008]; in particular the skill in providing regional climate change information in a given context needs to be evaluated as thoroughly as possible.

For a given emission scenario, the skill of regional climate change projections is limited by uncertainties mainly due to model errors and internal climate variability [*Stainforth et al.*, 2007]. In the context of downscaling, the question of skill can therefore be broken down into three sub-questions:

1. How well do GCMs simulate the input for regional climate change projections?
2. How well do downscaling methods work, in particular under climate change?
3. How strong is the signal-to-noise ratio between climate change trends and internal climate variability at regional scales?

In response to the first of these sub-questions, *Déqué et al.* [2007] showed that the GCM error is an important source of uncertainty in regional climate projections. For instance, European climate is affected by dynamical processes such as the polar jet stream, the North Atlantic storm tracks, stationary planetary waves or sudden stratospheric warmings; all are represented in GCMs with fairly high uncertainties [*Woollings*, 2013]. Some biases are common to the majority of GCMs and thereby limit the usefulness of even the most comprehensive model ensembles. A prominent example is the North Atlantic sea surface temperature cold bias that in turn causes a biased response of the atmospheric circulation [*Keeley et al.*, 2012].

Second, the downscaling method itself is often a considerable source of uncertainty [*Maraun et al.*, 2010; *Casanueva et al.*, 2013]. RCMs represent sub-grid processes by parameterizations—semi-empirical models that are tuned to best represent typical present day weather conditions across a whole model domain. Statistical downscaling methods represent complex scale interactions by relatively simple empirical models. How accurately downscaling methods under these limitations can capture climate change is a matter of current research. For instance, recent research suggests that, e.g., convection parameterizations might not correctly represent the response of extreme convective precipitation to climate change [*Kendon et al.*, 2014].

Finally, internal climate variability has recently been identified as a major source of uncertainty of regional climate projections [*Hawkins and Sutton*, 2009; *Deser et al.*, 2012; *Maraun*, 2013]. Any externally forced climate response is superimposed by internally generated climate fluctuations. The amplitude of these fluctuations depends on the variable, the region, season, temporal, and spatial scale considered. In contrast to model errors, uncertainties due to internal variability are fundamentally irreducible beyond the timescales at which they are predictable. Thus, even if a hypothetically perfect climate model were available, climate projections for lead times of several decades could be dominated by random fluctuations rather than climate change trends, especially at regional scales.

As spread in regional climate simulations caused by model uncertainties on the one hand and internal climate variability on the other hand is fundamentally different, their different effects on the decision process in a specific context have to be precisely understood.

A comprehensive effort to assess the credibility of regional climate change scenarios has to address the three sources of uncertainty discussed above. Such an effort can build upon existing downscaling intercomparison projects such as STARDEX [*Goodess et al.*, 2010], ENSEMBLES [*van der Linden and Mitchell*, 2009], NARCCAP [*Mearns et al.*, 2009, 2012] or CORDEX [*Giorgi et al.*, 2009], that already provide a wealth of information and actual high resolution climate simulations. Yet additionally, the wealth of statistical downscaling methods developed by individual climatologists, hydrologists, and statisticians [*Maraun et al.*, 2010] has to be integrated and compared relative to each other as well as with dynamical downscaling methods. Furthermore, the validation should in particular consider aspects that have received limited attention so far, such as extreme events, spatial-temporal dependencies and inter-variable relationships. The design of the validation experiment, the choice of meteorological variables and the aspects to be validated should be guided by user requirements. For Europe, the European Union Cooperation in Science and Technology (EU COST) Action ES1102 VALUE (www.value-cost.eu) attempts such an effort. COST Actions are funded networks that aim to coordinate existing research. They are therefore the ideal tool to develop common standards and to foster scientific dialogue.

VALUE addresses robust adaptation, GCM errors, and internal climate variability in a series of workshops. Yet the main focus of VALUE is to develop a common framework for the validation of downscaling methods. VALUE carries out its work in close interaction with users and thus helps to actively transfer scientific knowledge to stakeholders. In this paper, we present the validation framework designed by VALUE over the last two years. Starting with a scientific workshop in March 2012, statistical and dynamical downscalers together with statisticians as well as hydrologists, forestry, and agricultural scientists, representatives of environment agencies and international authorities have been involved in the design of the research agenda and the framework development. In the following, we refer to all latter users of downscaling products, whether they are scientists interested in modelling climate change impacts or decision makers having to consider climate change, simply as users.

2. Rationale and Overview

The climate system is complex and high-dimensional, and models are not intended to be isomorphisms of nature [*Stainforth et al.*, 2007]; thus no climate model or downscaling method can be expected to reproduce all aspects of the system perfectly, and a validation of all aspects would be practically impossible. However, in any given application only a small part of the system will be relevant: specific variables or phenomena, at specific space and time scales in a specific region. A user focused approach to validation must therefore start by identifying the phenomena and scales of interest; with respect to these, it must seek to identify the key strengths and weaknesses of a method. For a given application, it has to give advice whether a method performs well or even better than other methods, and where it is likely to fail. In this way, users can determine whether a particular method is appropriate for their application, and can compare methods.

The details will be application-dependent in any such user-focused approach to validation. Yet some general requirements can be formulated for a comprehensive user-focused intercomparison of downscaling methods for climate change studies. Thus the validation framework should (1) be transparent and provide relevant and defensible guidance for users; (2) assess the performance of the method under climate change as far as possible; (3) allow, in principle, for a comparison of all different types of dynamical and statistical downscaling approaches.

The first requirement implies the calculations involved in the exercise to be as simple as possible, with a clear documentation of the considered model and the validation procedure. Furthermore, it demands for readily available and high quality observational reference data. The second requirement implies an assessment whether a method correctly captures long-term variability, in particular the response to changes in external forcing. The third requirement has basically two implications: not all phenomena simulated by RCMs can be validated; the validation is restricted to phenomena that are represented by typical statistical downscaling models. Thus, although very important for enhancing our understanding and developing improved climate models, the validation of specific physical processes within RCMs is not part of the VALUE validation framework. Furthermore, it restricts the choice of performance measures. For instance, many stochastic downscaling methods provide time-varying probability density functions (pdfs) of a local predictand; these pdfs could be validated using sophisticated measures that have been developed to assess the skill of probabilistic weather forecasts [*Jolliffe and Stephenson*, 2003]. However, these measures are not easily applicable to RCMs and deterministic statistical downscaling methods, which provide a single downscaled field or sequence for a given GCM run. Yet stochastic methods allow for random sampling of multiple time series, which can be handled as an ensemble of deterministically downscaled fields. Thus, the framework relies solely on measures that are applicable to deterministic output.

The validation framework is intended to serve two purposes. First, VALUE will implement it to carry out a comprehensive intercomparison study. This exercise will be open to the scientific community—every developer and user of a downscaling method can contribute to the study by following the procedure laid out below and uploading the downscaled results to the VALUE webpage. Yet independent of our specific implementation, the VALUE framework is intended as a guideline for other validation studies, e.g., in regions and for variables different to those considered by VALUE. As the complete framework is rather complex, the actual implementation is separated into three tiers that will be carried out successively. Tier I comprises the most essential experiments and aspects of the validation that should be considered in any validation exercise. Tier II will address sub-daily time scales as well as spatial dependence and inter-variable aspects. These may be required in more specialized applications, and not all downscaling methods are designed to reproduce them. Additionally, tier II will cover the validation in a pseudo reality: future climate model simulations will be used as a testbed for downscaling methods. Tier III will assess the overall performance including GCM errors.

Core to our user-focused validation approach is a validation tree to select appropriate validation indices. It will be explained in Section 3. The validation itself is organized in specific experiments to isolate different aspects of the representation of regional climate. These experiments and their implementation will be laid out in Section 4. The reference observations cover both station and gridded data across different European climates. Daily data are complemented by a selection of sub-daily data to validate sub-daily downscaling methods. Pseudo reality data will be used to assess the performance of statistical methods in different climates. The specific data sets used for the validation will be presented in Section 5.

3. Validation Tree

A validation ultimately consists of deriving climate indices from model output, comparing these indices to reference indices calculated from observational data and quantifying the mismatch with the help of suitable performance measures (we use the term “index” in a very general way, including not only single numbers but also vectors such as time series). Often, however, aspects of a downscaling method have been validated that are of only little relevance for the problem to be addressed by the method (for instance, the 95th percentile of daily precipitation has frequently been used as an index for extreme precipitation, even though the corresponding events occur every 20 wet days).

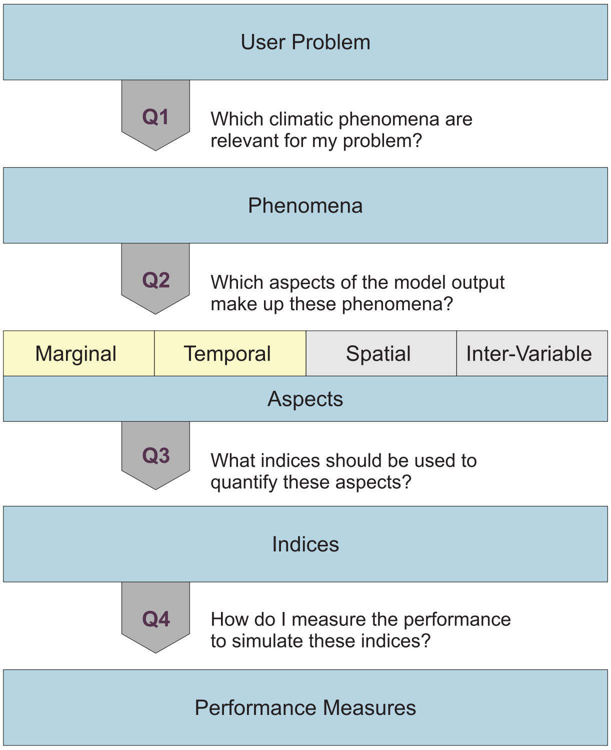
User Problem
Q1
Which climatic phenomena are relevant for my problem?
Phenomena
Q2
Which aspects of the model output make up these phenomena?
Marginal
Temporal
Spatial
Inter-Variable
Aspects
Q3
What indices should be used to quantify these aspects?
Indices
Q4
How do I measure the performance to simulate these indices?
Performance Measures

Figure 1. Validation Tree. Grey arrows: selection questions. Beige: tier I aspects; gray: tier II aspects.

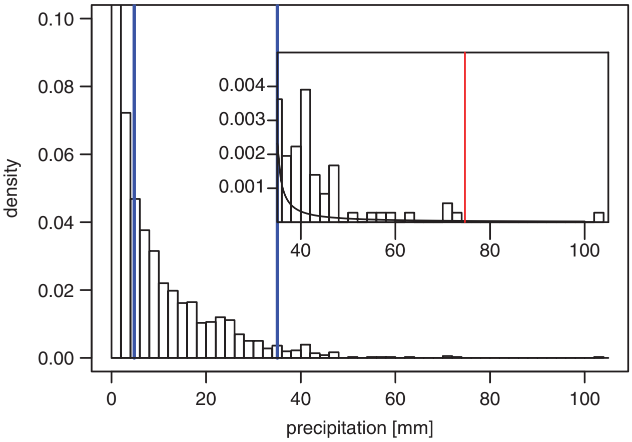

Figure 2. Marginal aspects. Histogram of observed summer daily precipitation at a grid-box in Norway, at 60.18°N and 5.52°E. Blue lines: empirical median and 98th percentile for wet days. The subpanel magnifies values beyond the 98th percentile. Black line: Generalized Pareto (GP) distribution for exceedances of 98th percentile. Red line: 20 summer return level based on GP distribution.

To guide the selection of relevant indices for different user problems, VALUE has therefore developed a decision tree for the selection of relevant indices and performance measures (Figure 1). From a user perspective, one would start by choosing *phenomena* relevant for the problem under consideration. Phenomena can be, e.g., extreme rainfall events, growing season or heat waves. Note that these phenomena might be compound. In the next step, one would consider the climate system as a multivariate distribution having marginal, temporal, spatial, and inter-variable *aspects* (for explanations and examples, see the end of this section). One would then ask which of these aspects are involved in the considered phenomena. For instance, the validation of extremes might involve marginal aspects such as the intensity, temporal aspects such as the seasonality or the length of extreme spells, or spatial aspects such as the spatial extent of an event. In the third step, one would select *indices* to quantify the considered aspects. Finally one would select appropriate *performance measures*, such as biases or skill scores, to compare indices derived from model data with those from observations.

As example for a user problem consider the construction of a new hydroelectric dam. To inform the design of the dam about potential future climatic conditions, downscaled climate model output could be used to drive a hydrological model that simulates the inflow into the reservoir. For a reliable estimate of the average power potential provided by the dam, overall precipitation is the meteorological phenomenon of interest. To avoid overflow of the reservoir, not only mean precipitation but also extreme precipitation events and their antecedent conditions are relevant. Finally, the risk of the reservoir falling dry is relevant, i.e., the risk of drought.

Consequently, relevant marginal aspects are average and extreme precipitation. Figure 2 shows the empirical marginal distribution of observed daily summer (June–August) precipitation at a grid box in Norway, at 60.18°N and 5.52°E (based on the gridded data set presented in Section 5). The left blue line indicates

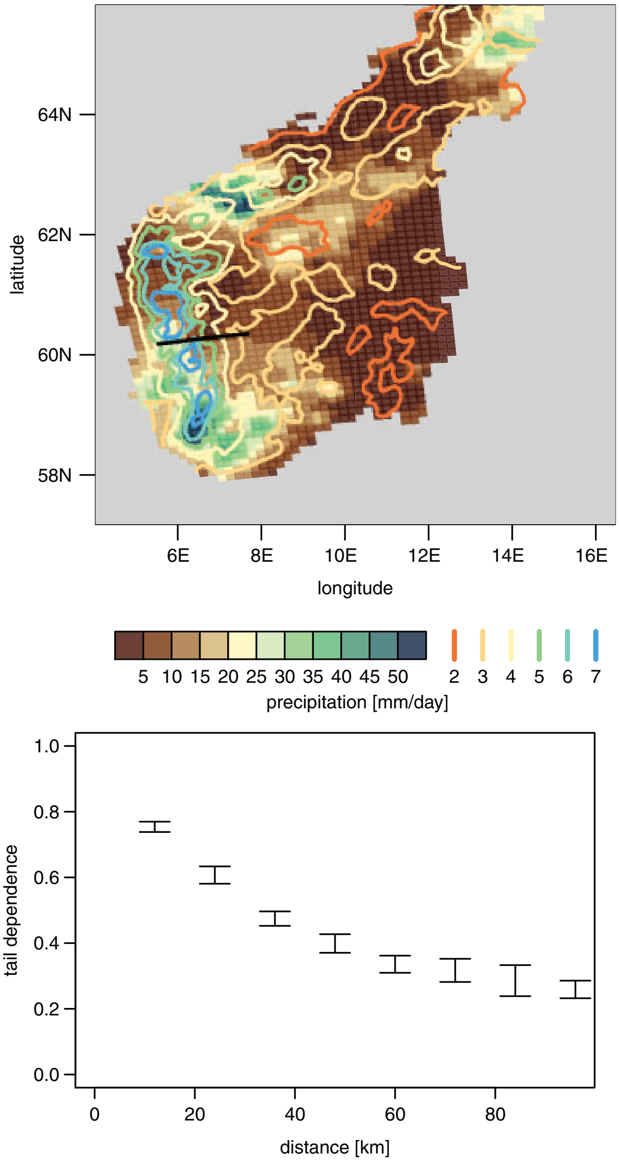

Figure 3. Spatial aspects. Top panel: filled grid boxes: total precipitation on 23 June 2000. Contour lines: climatological average daily summer precipitation (1980–2010). Bottom panel: empirical tail dependence coefficient as function of grid-box distance along the black line in left panel.

the median of the empirical distribution, a typical index measuring the core of the distribution. The right blue line indicates the empirical 98th percentile, a measure for heavy precipitation. The subpanel magnifies the distribution beyond the 98th percentile. The black line is the probability density derived from a Generalized Pareto distribution, fitted to all data beyond the 98th percentile. As an index for extreme precipitation, the red line finally indicates the 20-summer return level calculated from the Generalized Pareto distribution.

The temporal aspect of drought can be measured by indices characterizing dry spells. Long, but not extreme dry spells could, e.g., be characterized by the 90th percentile of the distribution.

Depending on the catchment characteristics and size [*Segond et al.*, 2007], spatial aspects might also be relevant in the example outlined above. An extreme flow event might be caused by heavy rainfall over a small area, but as well by widespread precipitation over the entire catchment. Thus, not only extremes at individual locations (=marginal aspects) are relevant, but also the risk of simultaneously high precipitation at different locations. To illustrate this point, consider Figure 2. The left panel depicts an individual event, namely daily precipitation on 23 June 2000 (filled grid boxes), overlaid by the climatological daily average rainfall (reference period 1980–2010, colored contour lines). Typically rainfall in Southern Norway, even during summer, is dominated by large scale precipitation along with the westerly flow. On average, therefore, precipitation is highest in the mountainous regions along the west coast, and lower in the rain shadow towards the hilly regions in the east. We refer to these climatological spatial patterns as systematic spatial variations (their characterization is discussed in Section 4.6).

During the event depicted in the top panel of Figure 3, a low pressure system over the North Sea moved moist air into Southern Norway, causing heavy precipitation in particular along the South coast and even further inland. In other words, precipitation (and other climatic processes) at different locations is not independent, but tends to co-vary across space. We refer to these co-varying anomalies about the climatological patterns as residual spatial dependence. They make up the spatial aspects. An index to measure the residual spatial dependence with a focus on extreme events could be based on the tail dependence coefficient. This index measures the probability that a high threshold is exceeded, given that a threshold at another location (or in another variable/at another time step, etc.) is exceeded simultaneously. The bottom panel of Figure 3 shows the empirical tail dependence of daily summer precipitation as a function

Table 1. List of Example Indices[a]

Aspect	Index	Performance Measure
Marginal	mean	bias/relative error
	variance	relative error
	20 season/year return level	bias/relative error
	number of threshold exceedances	bias
Temporal	time series	mean squared error/ correlation
	ACF lag 1, 2, 3	N.A.
	median of spell length distribution	bias
	90th percentile of spell length distrib.	bias
	minimum/maximum of annual cycle	bias/relative error
Spatial	decorrelation length	relative error
	variogram range	relative error
	decay length of tail dependence	relative error
Multivariate	Pearson/rank correlation	N.A.
	probability of joint exceedances	N.A.
	indices conditional on (no) exceedance	as above

[a] The complete list of indices may be found on www.value-cost.eu/indices. In some cases, model and observational indices will be affected by high uncertainties. Here no performance measures will be calculated (N.A. in the table), but just the index values will be given.

of grid-box distance along the black line shown in the top panel. As threshold, the 98th percentile of daily precipitation has been chosen individually for each grid box; the estimate for a specific distance is based on all grid box pairs of that distance. A derived index could, e.g., be the distance at which the tail dependence decays to $1/e$. Equally, one could investigate the decay in North–South direction.

Additionally, one might be interested in multivariate aspects. We interpret this aspect rather broadly, including local relationships between different predictand variables, but also teleconnection relationships between large-scale variability and local predictand variables. In the example, one might be interested in the relationship between temperature and precipitation to investigate whether precipitation correctly falls as snow or rain. Additionally, one might be interested whether the relationship of local precipitation with typical precipitation-causing weather patterns is correctly modelled, such as with the winter or summer North Atlantic Oscillation. In statistical downscaling, the latter relationships would ideally be accounted for by a careful selection of large-scale predictors. In practice, however, the response might not be correctly represented. This might also hold for RCMs that are not tuned to correctly simulate such relationships.

Finally, the performance of a model to represent these indices would then be measured typically by biases or relative errors between simulated and observed indices. The list of indices and corresponding performance measures is available from the VALUE website (www.value-cost.eu/indices) and will be continually updated. For an illustrative subset, see Table 1. In tier I of VALUE, only marginal and temporal aspects of the downscaled output will be considered. In tier II, the validation will be extended to spatial and inter-variable aspects.

4. Experimental Setup

Different experiments have been designed to identify different problems that might occur in the downscaling procedure: what is the isolated downscaling skill? How do statistical and dynamical methods compare? How do methods perform at different spatial scales? Do methods fail in representing regional climate change? How is the overall representation of regional climate, including errors inherited from GCMs? The experiments therefore differ in the type of predictor data and boundary conditions, the predictand data, and the space and time scales considered. For each applicable experiment, contributors will select a set of predictors or boundary conditions as input to their downscaling method, and upload the downscaled results to the VALUE website for a centralised validation. For statistical methods, the downscaled results for validation will be generated according to a prescribed cross validation procedure.

4.1. Experiment 1: Perfect Predictor

This is a standard experiment to validate the isolated downscaling skill regardless of errors in the large scale predictors or boundary conditions. Predictors are supposed to perfectly represent real weather at the synoptic scale. In practice, of course, they are taken from reanalysis data, which themselves are affected by errors due to limitations and changes in the underlying observational network, the assimilation technique and the employed general circulation models [*Brands et al.*, 2012]. RCM grid box output in general represents area averages; differences compared to station observations therefore result not only from model errors, but also from the scale gap between grid box and point scale. The latter discrepancy—which is not a model error—is known as the representativeness problem [*Klein Tank et al.*, 2009; *Zwiers et al.*, 2011].

4.1.1. Station Data (Tier I + II)

The aim of this experiment is to specifically test how well downscaling methods are able to represent point data, i.e., it will provide an overall assessment of model error and representativeness problem. The experiment will use reference data from a selection of at least 85 weather stations representative of the different climates and a variety of local characteristics in Europe. These data are selected based on expert judgment from the different countries accounting for representativeness, completeness, and the provision of different variables. In tier I, only marginal aspects will be considered, whereas in tier II marginal aspects along with inter-variable relationships will be considered.

4.1.2. Gridded Data (Tier I + II)

To isolate the downscaling performance from the representativeness problem, we will consider predictand data that are (re-)gridded to the same resolution as the considered RCM output. For statistical downscaling methods, the experiment will be carried out—depending on their purpose—across the complete regions (if the aim is to provide an alternative to RCMs across a large domain) or on a selection of grid boxes corresponding to a subset of the 85 selected stations (if the aim is to provide an alternative to RCMs for a small domain only). In tier I, only marginal aspects will be considered, in tier II also large-scale residual spatial dependence and inter-variable relationships.

4.1.3. Nested Station Data (Tier II)

This experiment is a variant of experiment 1(a) to specifically test the performance to simulate residual spatial dependence at different spatial scales ranging from the size of a whole country down to small regions. It is, e.g., conceivable that RCMs could outperform statistical methods in simulating residual spatial dependence at large scales, but at small scales could fall behind spatial statistical models that are specifically calibrated for a considered topography. This experiment will be carried out on example regions with a high station density.

4.1.4. Sub-Daily Data (Tier II)

This experiment is similar to experiment (a), but with the aim of testing downscaling methods at the sub-daily scale.

4.2. Experiment 2: Pseudo Reality (Tier II)

The ultimate goal of many downscaling studies is to downscale climate change simulations. Yet validating the skill of a downscaling method under climate change is difficult. Long series of high-quality observational reference data, as required to characterize past trends reliably, are scarce. Equally important, the availability of predictor data sets that cover a sufficiently long time period without temporal inhomogeneities, is also limited. As an additional test we therefore apply a pseudo reality, often called perfect model approach [*Charles et al.*, 1999; *Frías et al.*, 2006; *Vrac et al.*, 2007; *Maraun*, 2012; *Räisänen* and *Räty*, 2013]. A simulation of present and future climate with a specific GCM, downscaled with a specific RCM, is considered as pseudo reality—a testbed to identify potential problems of downscaling methods. For a given pseudo reality, different experiments are possible to test for different types of problems. In each case, the “pseudo-observed” predictands are taken from the RCM used in the pseudo reality.

4.2.1. Perfect Predictor (Same GCM/Same RCM)

This variant is designed to assess whether statistical downscaling and correction methods that use predictors from a GCM capture the climate change signal given perfect information on future predictors. To this end the predictors are taken from the same GCM that has been used to construct the pseudo reality. This experiment can reveal whether predictors necessary to simulate the response to climate change are missing, and whether the statistical model structure is unsuitable for extrapolations towards unobserved state space regions.

4.2.2. Imperfect Predictor (Same GCM/Different RCM)

This variant is designed to test whether an RCM bias correction method is able to correct RCM errors. To separate RCM errors, the predictors are taken from an RCM that is driven by the same GCM that has been used to construct the pseudo reality. The remaining RCM errors are mostly localized and stem from problems with parameterizations and surface boundary conditions such as orography [*Eden et al.*, 2012].

4.2.3. Imperfect Predictor (Different GCM/Same RCM)

This variant tests whether a correction of GCM biases, i.e., errors in the large scale boundary conditions is possible. To this end, the predictors are chosen from the same RCM that has been used to construct the pseudo reality, but driven with a different GCM. This test assesses the fundamental problems of GCM bias correction rather than the performance of individual bias correction methods. It will therefore be carried out with a limited number of example bias correction methods.

Other experiments are of course conceivable, e.g., an imperfect predictor experiment for statistical downscaling methods with GCM predictors, and similarly an imperfect predictor experiment for RCM bias correction methods with different GCM and RCM as in *Räisänen and Räty* [2013]. These experiments would assess the most realistic situation, yet they would not test downscaling model performance but rather the spread of GCMs. Therefore we will not consider these experiments.

4.3. Experiment 3: GCM Predictor (Tier III)

As discussed above, large errors in regional climate change projections are often inherited from the driving GCM. Therefore, validating only the downscaling performance is not sufficient to test the credibility of regional climate change scenarios. As a consequence, VALUE will also validate the combined GCM and downscaling performance. To avoid representativeness problems, gridded data will be taken as observational reference. The value of this experiment, however, is limited by internal climate variability. Differences between model output and observations—also on climatological time scales—will in general be a superposition of systematic model biases and a substantial contribution of internal variability.

4.4. Cross Validation

In order to avoid artificial skill, statistical models have to be validated on data that have not been used for calibration. For statistical downscaling models in experiment 1, a five-fold cross validation will therefore be applied: the chosen time period (1979–2008) will be divided into five non-overlapping blocks. In turn, the statistical models will be calibrated against four of these blocks and used to predict the remaining block. In total this yields one cross-validated model prediction. In case of pseudo reality experiments, the calibration period will cover simulated present day climate, whereas the validation period will cover simulated future climate. For experiment 3, no cross validation will be carried out: here it is essential to minimize the influence of internal climate variability on the validation results; therefore the selection of very long calibration and validation periods is more important than the elimination of artificial skill.

In general, the development of a statistical model requires a thorough selection of the model structure based on statistical measures, i.e., a selection of predictors and how these act upon the model parameters [*Davison*, 2003; *Maraun et al.*, 2010]. Strictly speaking, this selection would have to be carried out for each block. To reduce the computational burden, in particular for complex statistical models, we suggest to derive the model structure once from the whole data set, and then apply the cross validation keeping the model structure identical for each block. Of course, for dynamical downscaling no cross validation is required.

4.5. Upload and Validation

Each contributor is free to select the experiments to carry out, and the variables and time scales to provide. For each chosen experiment, grid box or station, season, and variable a contributor will upload downscaled time series to a central data server. In case of stochastic downscaling methods, 100 realizations have to be uploaded to reconstruct the resulting downscaled distribution. Indices and performance measures will be calculated in a centralized manner in order to avoid inconsistencies of program codes. The selection of validation indices will be determined by the meta-data provided by the contributor to avoid meaningless or misleading evaluations. Additionally, contributors have the option to comment on the validation results. The portal will offer three modes: a private mode to check the validation results, a VALUE mode to share the results internally for further research, and a public mode. Every upload will be logged for version tracking. The VALUE mode is designed for cases where the validation results are bad because of obvious and easily correctable shortcomings of a method, but where these shortcomings are of scientific interest. For instance, important predictors might be missing in a model version. Understanding the predictor influence might be of scientific interest, but the contributor might not want the model to be applied in any real-world applications, or might not want the model output to be misinterpreted. Details of data policies will be formulated in detail and made available on the VALUE portal. VALUE deliberately chose a grass roots approach not to be limited to a few selected downscaling methods, but to be as open as possible to the whole downscaling community. VALUE is aware of the fact, that misconduct in such a design cannot be fully excluded.

4.6. Presentation of Results and Summary Statistics

If a method has not been validated for a particular index or region, a default “N.V.” will be set to indicate that the method has not been validated, and therefore the developer of the method does not recommend to use the method in that context. This default setting can in particular be used to avoid an unnecessary validation for aspects a method is not designed to correctly simulate (e.g., multi-site aspects in case of a single-site method).

Some users might be interested in the performance for a specific grid box or station only, others will prefer to focus on the performance over a larger region. Therefore, spatial summary performance measures for each index will be considered, such as spatial root mean squared errors, mean biases, ratios of standard deviations, and pattern correlations. Also the local performance measures themselves will be averaged in space.

5. Validation Data

To enable a fair comparison that reveals relative strengths and weaknesses of different methods, the selected validation data span different climatic regions across Europe and consist both of station and gridded data. The default temporal resolution is daily, but sets of sub-daily data have been selected to validate sub-daily downscaling methods. To keep the exercise as open as possible, we restricted the selection to data that could be made publicly available either on the VALUE website itself, or—after registration—on the webpage of the respective weather service. The only exception are observations from Switzerland; these data are only available for use by VALUE members, but have nevertheless been selected because they are currently the only dense and reliable gridded data set covering an Alpine area in Europe. Finally, transient RCM data will be used as pseudo realities to assess the performance of statistical methods in different climates. The different regions covered by the observational data sets are shown in Figure 4. The validation domains and choice of data might be subject to future changes depending on the availability of additional reference data and quality issues that might occur.

Predictor data will be taken from ERA-Interim [*Dee et al.*, 2011], and the RCM ensemble participating in the validation will be taken from EURO-CORDEX [*Jacob et al.*, 2014]. The ERA-Interim driven experiments

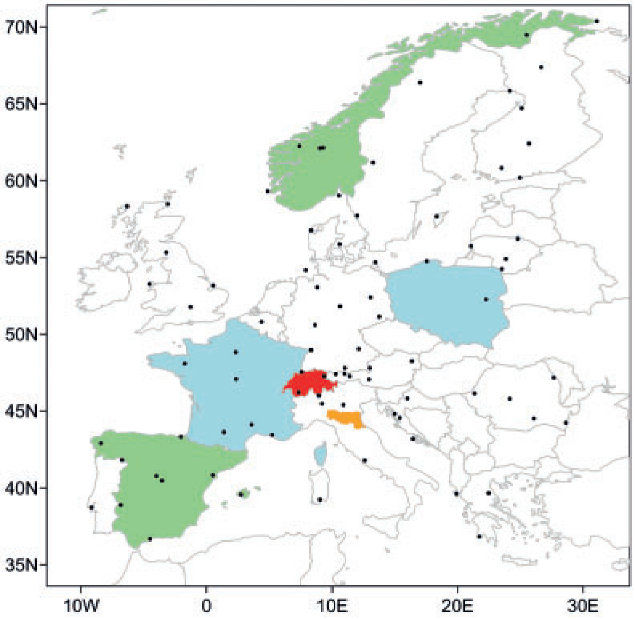

Figure 4. Validation Domains and Data. Green: gridded data without registration; blue: gridded data with registration; orange: sub-daily station data. Red: gridded data for use by VALUE members only. Additionally, publicly available daily station data across Europe will be selected. Black dots: selection of 85 stations.

carried out in EURO-CORDEX in general cover the period from 1989 to 2008, a subset of simulations the longer period from 1979 to 2008. To ensure as robust results as possible, the longer period was chosen as standard for VALUE experiments with observed predictands. The chosen predictand data and a selection of standard predictor data sets—or links to the data providers—can be found on www.value-cost.eu/data.

We note that all empirical data sources themselves might be subject to uncertainties. For instance, rain gage data might be affected by undercatch errors; gridding of station data is affected by sampling errors due to limited sub-grid data and the chosen aggregation method. VALUE has no resources to address these issues prior to the validation exercise—some might even be de facto unsolvable. In the course of the exercise, however, the importance of selected observational uncertainties will be studied.

5.1. Gridded Data

For validating downscaling methods in experiment 1(a), gridded observational data representing the spatial resolution of RCM output are required. The investigated RCM experiments will be provided by the high-resolution ERA-Interim driven RCM ensemble of the EURO-CORDEX initiative [*Kotlarski et al.*, 2014] carried out at a grid resolution of 0.11° on a rotated grid (approximately 12 km × 12 km). European-scale observational reference data on that spatial scale and at daily resolution are not available. Therefore, the VALUE validation exercise will focus on selected sub-domains of the European continent for which reliable high-resolution gridded temperature and precipitation data at daily resolution exist and if possible can be made publicly available. We require the underlying station network to be as dense as possible to keep uncertainties in the gridding procedure as low as possible. The selected sub-domains should sample the diversity of the European climate. Presently, gridded reference data at daily resolution have been obtained for Spain, France, Norway, Poland and—for VALUE members—Switzerland and have been (or will soon be) regridded to the 0.11° RCM resolution. These data sets will be provided for VALUE participants for calibration and validation purposes.

5.2. Station Data

The European Climate Assessment (ECA) dataset [*Klein Tank et al.*, 2002] provides series of daily observations at meteorological stations throughout Europe and the Mediterranean for a number of variables. Data are freely available for non-commercial research at thousands of stations for precipitation and temperature, with varying density across countries. The number of stations for alternative variables (e.g., wind and cloud cover) is low and restricted to a few countries. Only data for selected countries covering the reference ERA-Interim period 1979–2008 with less than 5% of missing values will be used. A sub-selection of currently 85 high-quality station time series has been chosen (see Figure 4).

5.3. Sub-Daily Data

The availability of reference data at sub-daily resolution is very limited. Validation exercises targeting quantities at sub-daily resolution will therefore only be carried out over a limited set of individual stations, potentially spread across Europe. Currently, sub-daily station data are available for Emilia-Romagna.

5.4. Pseudo Reality

For the pseudo-reality experiment 2, GCM-driven RCM simulations carried out within the EURO-CORDEX initiative at grid resolutions of approximately 12 and 50 km will be used [*Jacob et al.*, 2014]. These experiments represent multi-GCM – multi-RCM – multi-emission-scenario ensembles. In total, these ensembles will consist of about 45 experiments (12 km) and 70 experiments (50 km), respectively; about one-third of each ensemble is currently available via the Earth System Grid Federation archive (e.g., http://esgf-data.dkrz.de). For the purpose of VALUE, however, only a sub-ensemble will be used. A pseudo reality is not required to be a perfect copy of the real world; however, the simulated climate needs to be plausible in the sense that relevant processes are realistically simulated. This requirement limits the choice of indices for some predictand variables. For instance, the RCMs chosen as pseudo realities arguably underestimate the response of extreme summer convection to climate change [*Kendon et al.*, 2014], hence extreme precipitation will not be considered.

5.5. Predictor Data

Standard predictor data from ERA-Interim will be provided for the perfect predictor experiment 1 via the VALUE website. Contributors wishing to use predictors derived from other fields may do so; the only requirement is that the predictor data must be derived from the ERA-Interim reanalysis so as to guarantee that differences between downscaled fields are not due to the use of different reanalysis products. Model versions using different predictors count as different methods and can be validated for comparison.

6. Concluding Remarks

VALUE aims to provide a platform for the validation of regional climate scenarios and to foster networking between the different communities involved in the generation and application of these scenarios. VALUE is an open network and a community effort. We welcome contributions to VALUE from researchers and users across Europe and beyond. Please visit our webpage for advice how to join us. The VALUE framework is under continual development, and we are grateful for constructive comments and criticism.

The experiments carried out within VALUE will provide a comprehensive validation database of a wide range of state-of-the-art downscaling methods. For Europe, VALUE will give precise information about downscaling skill for a range of variables, for different aspects of these variables, for a representative selection of climates. End-users can enquire this data base to select suitable method and to quantitatively learn about limitations of these methods for a particular problem. The essential results of the experiments will be distilled in a series of scientific papers. Additionally it is anticipated to provide a guidance document written for users with different levels of scientific background knowledge. All in all, VALUE will provide urgently needed information for impact modelling and decision planning, no matter which stance one might assume in the debates highlighted in the introduction.

The framework presented in this paper has emerged from many wide-ranging discussions involving representatives from across Europe, with different backgrounds and priorities. During these discussions, many issues were raised that have not been incorporated into the framework described here, either because they were too complex to integrate them into a formalized framework to compare different types of downscaling approaches, because they seemed impracticable, because of limited resources, limited publicly available validation data, or simply because they were deemed irrelevant. For instance, the issue of validating the representation of physical processes in RCMs that are not simulated by typical statistical models has been discussed, and VALUE acknowledges the importance of such a validation. Yet it cannot be included in a generic framework that aims to compare dynamical and statistical downscaling. It has also been discussed to carry out a blind validation, i.e., providing reference observations without revealing the geographical location or other station identifier, and without providing the validation reference data. Yet the geographical location is in principle identifiable if only publicly available data sets are used; also a four-fold cross validation would be impossible without having a complete reference data set; finally, the geographical location is important for an educated predictor choice.

Even though not complete, VALUE is probably the most comprehensive effort to compare downscaling methods to date. With the experience gained over the last two years and the first results expected soon, we hope VALUE will make an important contribution to regional climate research. With this experience we

Acknowledgments
VALUE is funded via the EU FP7 programme as EU COST Action ES1102. We thank the VALUE community for their input to this framework and for coding the validation routines, in particular Judit Bartholy, Maria Casado, Paul Christodoulides, Jonathan Eden, Andreas Fischer, Adam Jaczewski, Malaak Kallache, Neyko Neykov, Christian Page, Valentina Pavan, Rita Pongracz, and Thordis Thorarinsdottir. We gratefully acknowledge the provision of observational reference data by MET Norway, IMGW, MeteoSwiss, Meteo France, ARPA Emilia-Romagna and AEMET/University of Cantabria. For details see www.value-cost.eu/data. The authors gratefully acknowledge helpful comments by two anonymous reviewers.

believe the VALUE framework can be integrated into and inspire other international validation exercises, such as the newly funded CORDEX-ESD.

References

Barsugli, J., et al. (2013), The practitioner's dilemma: How to assess the credibility of downscaled climate projections, *Eos Trans. AGU*, *94*(46), 424–425.

Brands, S., J. Gutiérrez, and S. Herrera (2012), On the use of reanalysis data for downscaling, *J. Clim.*, *25*, 2517–2526, doi: 10.1175/JCLI-D-11-00251.1.

Brown, C., Y. Ghile, M. Laverty, and K. Li (2012), Decision scaling: Linking bottom-up vulnerability analysis with climate projections in the water sector, *Water Resour. Res.*, *48*, W09537, doi:10.1029/2011WR011212.

Casanueva, A., S. Herrera, J. Fernández, M. Frías, and J. Gutiérrez (2013), Evaluation and projection of daily temperature percentiles from statistical and dynamical downscaling methods, *Nat. Hazards Earth Syst. Sci.*, *13*, 2089–2099, doi:10.5194/nhess-13-2089-2013.

Charles, S., B. Bates, P. Whetton, and J. Hughes (1999), Validation of downscaling models for changed climate conditions: Case study of southwestern Australia, *Clim. Res.*, *12*, 1–14.

Davison, A. C. (2003), *Statistical Models, Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge Univ. Press, Cambridge U. K.

Dee, D., et al. (2011), The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Q. J. R. Meteorol. Soc.*, *137*, 553–597.

Déqué, M., D. P. Rowell, D. Luthi, F. Giorgi, J. H. Christensen, B. Rockel, D. Jacob, E. Kjellström, M. de Castro, and B. van den Hurk (2007), An intercomparison of regional climate simulations for Europe: Assessing uncertainties in model projections, *Clim. Change*, *81*, 53–70.

Deser, C., R. Knutti, S. Solomon, and A. Phillips (2012), Communication of the role of natural variability in future North American climate, *Nat. Clim. Change*, *2*, 775–779.

Deser, F., B. Rockel, H. von Storch, J. Winterfeldt, and M. Zahn (2011), Regional climate models add value to global model data. a review and selected examples, *Bull. Am. Meteorol. Soc.*, *92*, 1181–1192.

Dessai, S. (2009), Do we need better predictions to adapt to a changing climate?, *Eos Trans. AGU*, *90*, 111–112., doi:10.1029/2009EO130003

Eden, J., M. Widmann, D. Grawe, and S. Rast (2012), Skill, correction, and downscaling of GCM-simulated precipitation, *J. Clim.*, *25*, 3970–3984.

European Commission (2013), An EU Strategy on adaptation to climate change, http://ec.europa.eu/clima/policies/adaptation.

Fowler, H. J., S. Blenkinsop, and C. Tebaldi (2007), Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling, *Int. J. Climatol.*, *27*, 1547–1578.

Frías, M., E. Zorita, J. Fernández, and C. Rodríguez-Puebla (2006), Testing statistical downscaling methods in simulated climates, *Geophys. Res. Lett.*, *33*, L19,807., doi:10.1029/2006GL027453

Giorgi, F., C. Jones, and G. Asrar (2009), Addressing climate information needs at the regional level: The CORDEX framework, *WMO Bull.*, *58*(3), 175–183.

Goodess, C., et al. (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, *Tech. rep.*, Climatic Research Unit.

Hawkins, E., and R. Sutton (2009), The potential to narrow uncertainty in regional climate predictions, *Bull. Am. Meteorol. Soc.*, *90*(8), 1095–1107.

Hewitson, B., J. Daron, R. Crane, M. Zermoglio, and C. Jack (2013), Interrogating empirical-statistical downscaling, *Clim. Change*, *122*, 539–554.

Jacob, D., et al. (2014), EURO-CORDEX: New high-resolution climate change projections for European impact research, *Reg. Environ. Change*, *14*, 563–578.

Jolliffe, I. T., and D. B. Stephenson (Eds) (2003), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Wiley, Chichester, U.K.

Keeley, S., R. Sutton, and L. Shaffrey (2012), The impact of North Atlantic sea surface temperature errors on the simulation of North Atlantic European region climate, *Q. J. R. Meteorol. Soc.*, *138*, 1774–1783.

Kendon, E., N. Roberts, H. Fowler, M. Roberts, S. Chan, and C. Senior (2014), Heavier summer downpours with climate change revealed by weather forecast resolution model, *Nat. Clim. Change*, *4*, 570–576.

Klein Tank, A., et al. (2002), Daily dataset of 20th-century surface air temperature and precipitation series for the european climate assessment, *Int. J. Climatol.*, *22*(12), 1441–1453.

Klein Tank, A., F. Zwiers, and X. Zhang (2009), Guidelines on analysis of extremes in a changing climate in support of informed decisions for adaptation, *Climate data and monitoring wcdmp-no. 72*, World Meteorological Organisation.

Kotlarski, S., et al. (2014), Regional climate modelling on European scales: A joint standard evaluation of the EURO-CORDEX RCM ensemble, *Geosci. Model Dev.*, *7*, 1297–1333.

Kundzewicz, Z., and E. Stakhiv (2010), Are climate models "ready for prime time" in water resources management applications, or is more research needed?, *Hydrol. Sci. J.*, *55*, 1085–1089.

Lempert, R., N. Nakicenovic, D. Sarewitz, and M. Schlesinger (2004), Characterising climate-change uncertainties for decision-makers, *Clim. Change*, *65*, 1–9.

Maraun, D. (2012), Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums, *Geophys. Res. Lett.*, *39*, L06,706., doi:10.1029/2012GL051210

Maraun, D. (2013), When will trends in European mean and heavy daily precipitation emerge?, *Environ. Res. Lett.*, *8*, 014,004.

Maraun, D., et al. (2010), Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user, *Rev. Geophys.*, *48*, RG3003, doi:10.1029/2009RG000314.

Mearns, L., W. Gutowski, R. Jones, L.-Y. Leung, S. McGinnis, A. Nunes, and Y. Qian (2009), A regional climate change assessment program for North America, *Eos Trans. AGU*, *90*(36), 311–312.

Mearns, L., et al. (2012), The North American Regional Climate Change Assessment Program. Overview of Phase I results, *Bull. Am. Meteorol. Soc.*, *93*, 1337–1362.

Prudhomme, C., R. Wilby, S. Crooks, A. Kay, and N. Reynard (2010), Scenario-neutral approach to climate change impact studies: Application to flood risk, *J. Hydrol.*, *390*, 198–209.

Räisänen, J., and O. Räty (2013), Projections of daily mean temperature variability in the future: cross-validation tests with ENSEMBLES regional climate simulations, *Clim. Dyn.*, *41*, 1553–1568.

Rummukainen, M. (2010), State-of-the-art with regional climate models, *Wiley Int. Rev. Clim. Change*, *1*, 82–96, doi:10.1002/wcc.8.

Segond, M.-L., H. Wheater, and C. Onof (2007), The significance of spatial rainfall representation for flood runoff estimation: A numerical evaluation based on the Lee catchment, *J. Hydrol.*, *347*, 116–131.

Stainforth, D., M. Allen, E. Tredger, and L. Smith (2007), Confidence, uncertainty and decision-support relevance in climate predictions, *Philos. Trans. R. Soc. A*, *365*, 2145–2161.

UNFCCC (2008), Compendium on methods and tools to evaluate impacts of, and vulnerability and adaptation to climate change.

United Nations (1992), United Nations Framework Convention on Climate Change.

van der Linden, P., and J. F. B. Mitchell (2009), ENSEMBLES: Climate change and its impacts: Summary of research and results from the ENSEMBLES project, *Tech. rep.*, Met Office Hadley Centre.

Vrac, M., M. L. Stein, K. Hayhoe, and X. Z. Liang (2007), A general method for validating statistical downscaling methods under future climate change, *Geophys. Res. Lett.*, *34*, L18,701, doi:10.1029/2007GL030295.

Wilby, R. (2010), Opinion: Evaluating climate model outputs for hydrological applications, *Hydrol. Sci. J.*, *55*, 1090–1093.

Wilby, R., and S. Dessai (2010), Robust adaptation to climate change, *Weather*, *65*, 180–185.

Woollings, T. (2013), Dynamical influences on European climate: An uncertain future, *Philos. Trans. R. Soc. A*, *368*, 3733–3756.

Worldbank Independent Evaluation Group (2012), Adapting to climate change: Assessing the World Bank Group Experience Phase III.

Zwiers, F., L. Alexander, G. Hegerl, T. Knutson, P. Naveau, N. Nicholls, C. Schär, S. Seneviratne, and X. Zhang (2011), Community paper on climate extremes. Challenges in estimating and understanding recent changes in the frequency and intensity of extreme climate and weather events.