A Hybrid Downscaling Approach for Future Temperature and Precipitation Change

HELENE BIRKELUND ERLANDSEN,^a KAISA M. PARDING,^a RASMUS BENESTAD,^a
ABDELKADER MEZGHANI,^a AND MARIE PONTOPPIDAN^b

^a Norwegian Meteorological Institute, Oslo, Norway; ^b Bjerknes Centre for Climate Research,
NORCE Norwegian Research Centre, Bergen, Norway

(Manuscript received 10 January 2020, in final form 7 July 2020)

ABSTRACT: We used empirical–statistical downscaling in a pseudoreality context, in which both large-scale predictors and small-scale predictands were based on climate model results. The large-scale conditions were taken from a global climate model, and the small-scale conditions were taken from dynamical downscaling of the same global model with a convection-permitting regional climate model covering southern Norway. This hybrid downscaling approach, a “perfect model”-type experiment, provided 120 years of data under the CMIP5 high-emission scenario. Ample calibration samples made rigorous testing possible, enabling us to evaluate the effect of empirical–statistical model configurations and predictor choices and to assess the stationarity of the statistical models by investigating their sensitivity to different calibration intervals. The skill of the statistical models was evaluated in terms of their ability to reproduce the interannual correlation and long-term trends in seasonal 2-m temperature T_{2m} , wet-day frequency f_w , and wet-day mean precipitation μ . We found that different 30-yr calibration intervals often resulted in differing statistical models, depending on the specific choice of years. The hybrid downscaling approach allowed us to emulate seasonal mean regional climate model output with a high spatial resolution (0.05° latitude and 0.1° longitude grid) for up to 100 GCM runs while circumventing the issue of short calibration time, and it provides a robust set of empirically downscaled GCM runs.

KEYWORDS: Principal components analysis; Climate prediction; Cloud resolving models; General circulation models; Climate services

1. Introduction

Global climate models (GCMs) are our main tool for providing plausible outlooks for what future climate conditions we can expect with changes in atmospheric greenhouse gas concentrations and other forcings (IPCC AR5; CMIP5). They are designed to capture large-scale phenomena and processes in the climate system; however, the amount of details they can provide is limited by design and computational resources. The raw output of the GCMs has served as a useful source of information for mitigation purposes (IPCC 2013). However, to provide information on local scales needed for impact studies and climate change adaptation, it is necessary to downscale the global models (Takayabu et al. 2015).

Two types of methods have been used to downscale GCM data: empirical–statistical downscaling (ESD) and dynamical downscaling. ESD involves using statistical models to transfer information from a large-scale predictor field, that the global models are able to reproduce, to compatible information on a fine spatial scale. In other words, ESD utilizes dependencies

between the spatial scales found in the data and applies them to GCM data to make projections on a smaller scale when only projections on a coarse scale are available. Dynamical downscaling involves running a regional climate model (RCM) with a finer resolution than the GCM, ingesting data from the GCM on its lateral and surface boundaries. The two downscaling methods complement each other as they have different strengths and weaknesses.

Dynamical downscaling has the value of being based on physical processes in the atmosphere and provides a description of a complete set of variables over a volume of the atmosphere. Nevertheless, assumptions and statistical relationships are also used in RCMs (see, e.g., Bellprat et al. 2012). The RCMs are more comprehensive than statistical downscaling in terms of output variables and offer both high temporal and spatial resolution. However, dynamically downscaled results of GCM often have climatological biases, either inherited from the driving GCM or originating from the RCM itself. Such biases may be due to configuration choices or drift in the model (Kotlarski et al. 2017; Fernández et al. 2019; Pontoppidan et al. 2019). Dynamical downscaling is also computationally costly and typically cannot be applied to a very large ensemble of GCMs. The choice of RCM and its configuration options, for example, domain size, resolution, inclusion of spectral nudging, choice of physical schemes, implementation of time varying climate gases and aerosols, or land-use change, as implemented

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JAMC-D-20-0013.s1>.

Corresponding author: Helene Birkelund Erlandsen, helenebe@met.no

DOI: 10.1175/JAMC-D-20-0013.1

© 2020 American Meteorological Society

This article is licensed under a [Creative Commons Attribution 4.0 license](http://creativecommons.org/licenses/by/4.0/) (<http://creativecommons.org/licenses/by/4.0/>).

in the IPCC scenarios, results in a span of possible downscaled results based on a single GCM run. It is also possible, and not necessarily physically incorrect, that the downscaled results show a different local trend than the driving GCM model (Bartók et al. 2017; Enriquez-Alonso et al. 2017).

As with dynamical downscaling, ESD has both strengths and limitations, the major strength being that it brings together model results and observational data. The use of statistical models requires a good understanding of the physical processes that one tries to model, statistical theory, and the data themselves as unsuited methods can lead to misleading results. It is crucial that the ESD model is thoroughly evaluated before being used to make projections of the local climate. On the other hand, if appropriately framed, a suitable choice of statistical methods may provide valuable and reliable information, since the statistical characteristics of the climate system often are highly predictable. Moreover, ESD is computationally efficient and thus has the capability to include information from a full ensemble of GCM runs. Its efficiency makes it possible to downscale large ensembles of GCM runs and enables capturing the internal variability of the GCMs, which has a considerable influence on regional scales (Deser et al. 2012; Benestad et al. 2016). The number of GCMs included in a multimodel ensemble has been shown to significantly impact the representation of climate change (Mezghani et al. 2019), and the difference between the mean climate change of a small set of GCMs can be as large as the climate change itself. ESD is, however, limited by the availability of observations that may lack both spatial and temporal coverage, and it may contain errors. A likely minimum amount of years of observational data required to calibrate a statistical model is 30–45 years, depending on the character of the regional climate. Long records of high-quality observational data are scarce, but Europe is exceptional in terms of data availability with ~70 years of observations and high-resolution gridded observations (Cornes et al. 2018). An open question is whether a statistical model calibrated with data from one period is applicable to a different period with a changed climate. This is referred to as the question of “model stationarity” (Wilby 1997; Schmith 2008; Dixon et al. 2016). For ESD methods based on linear regression, the assumption of model stationarity means that transfer coefficients [β_{ij} in Eq. (2), below] are expected to be independent of the calibration period, and if this is not the case, then the statistical model and projections based on it are not robust.

The effect of local feedback mechanisms with a nonlinear dependence on climate change may not be accurately represented in a statistical model and may pose a problem to ESD (Walton et al. 2017; Lanzante et al. 2018). One example is the temperature threshold dependency of local snow cover or vegetation, which again influences other variables, particularly local near surface temperature. However, the timeliness of these local tipping points may not be well represented in a dynamical downscaling either (see, e.g., Fernández et al. 2019).

A pioneering work using a “perfect model” experiment and RCM data is described in Charles et al. (1999), in which RCM fields were used as both fine-scale predictand data and

coarse-scale predictor data. A statistical model was trained to emulate precipitation fields from an RCM run under a $1 \times \text{CO}_2$ scenario using RCM fields regridded to a coarser resolution as large-scale predictors. The model was subsequently applied using the coarsened RCM fields from a $2 \times \text{CO}_2$ scenario run as predictors and validated on $2 \times \text{CO}_2$ scenario RCM precipitation. In Vrac et al. (2007), the question of “model stationarity” was investigated by validating historically calibrated ESD methods on the RCM output for a future time period using the RCM output as pseudo-observations. However, they did not make use of the future RCM output for calibration, and the RCM output covered only two 10-yr time slices (1990–99 and 2090–99). Vrac et al. (2007) stated that longer time periods would provide more robust statistical relationships and that 10 years should be viewed as a minimum requirement.

Of recent studies that have scrutinized the underlying assumption of stationarity in statistical downscaling models under climate change, two have involved “perfect model” experiments by using high-resolution GCM data and coarsened GCM data to test the sensitivity of the distributional mapping approach to climate change (Dixon et al. 2016; Lanzante et al. 2018). Parding et al. (2019) used an empirical–statistical method to obtain projections of the seasonal North Atlantic cyclone density from seasonal GCM fields and assessed the stationarity of the statistical models by varying the calibration period and comparing the predictions to the original high frequency fields. The results showed that the statistical models based on 500-hPa geopotential height were highly dependent on the calibration period, meaning this was not a robust predictor.

We here use the term “hybrid downscaling” when referring to downscaling based on both dynamical and empirical statistical approaches, where the dynamical downscaling results are used as predictands in statistical model calibration. Recent examples of hybrid statistical–dynamical downscaling approaches are Li et al. (2012); Berg et al. (2015); Sun et al. (2015); Walton et al. (2015, 2017). In Li et al. (2012), available CMIP3 GCM data and dynamically downscaled results from the North American Regional Climate Change Assessment Program at a 45-km horizontal resolution were used to construct and evaluate linear regression models emulating the GCM–RCM relationships, using both an historic and a future 30-yr period for calibration and evaluation. Separate calibration on each of the two 30-yr intervals indicated nonlinearity in RCM response, justifying a time-trend component in the linear regression model. In the latter four studies the future climatological changes in monthly mean temperature (Walton et al. 2015, 2017) and precipitation (Berg et al. 2015) in California were emulated by using dynamically downscaled output from five GCMs to construct ESD models generating pseudo-RCM changes for full ensembles of GCMs. In this case, the dynamical downscaling adopted a “pseudo-global-warming” approach, that is, the RCM driving data were historical reanalysis data with GCM specific perturbations added according to the GCM’s climatological change in the relevant variables and domain. The pseudo-global-warming runs covered limited time slices of the future: 10-yr future time-slice runs in Walton et al. (2017)

FIG. 1. End-of-century changes in seasonal mean (left) 2-m temperature T_{2m} , (center) wet-day mean precipitation μ , and (right) the number of rainy days per season n_w for the dynamically downscaled fields of the climatologically bias-corrected NorESM1-M r111p1. The change was estimated as the linear trend for the 1980–2099 period multiplied by the length of the period (120 years). The black shading indicates where the trend cannot be distinguished from noise ($p > 0.025$ in a two-sided Student's t test or equivalently CI $< 95\%$). The areal means are displayed in the upper-left corner of each plot.

and only 3-yr time slices for four GCMs and 20 years for a single GCM in Berg et al. (2015) and Walton et al. (2015).

To our knowledge, there have been few attempts to construct perfect model experiments to test the ESD framework based on GCM and RCM data paired via dynamical downscaling covering several decades, let alone more than a century, of climate evolution. In this study, we had access to dynamically downscaled GCM data from 1980 to 2100 assuming the high-emission scenario [representative concentration pathway (RCP) 8.5]. The availability of 120 consecutive years under a period of pronounced climate change presented an opportunity to systematically address the question of model stationarity when constructing ESD-based projections. In this study, we investigated the sensitivity of ESD-derived seasonal 2-m surface temperature T_{2m} , the wet-day mean precipitation μ , and the wet-day frequency f_w to varying experimental setups and calibration times, and predictor variable choice. We downscale seasonal aggregates of daily temperature and precipitation rather than daily data itself because 1) the statistical properties tend to be more predictable than single outcomes and 2) it is more computationally efficient. We used μ and f_w to characterize precipitation as these two key parameters can be used to describe the statistical distribution of mean precipitation, and the likelihood of extreme precipitation (Benestad et al. 2019). The impact of using bias-corrected GCM fields was also investigated.

Our main working hypotheses were that

- 1) using GCM fields for which the mean has been climatologically bias corrected as predictors does not improve the ESD estimates,
- 2) detrending the data prior to calibration facilitates a more robust model calibration, and
- 3) a period of 30 years is sufficient for skillful model calibration.

The veracity of these hypotheses was assessed in terms of the skill of the ESD results. The most skillful ESD designs were subsequently applied to many GCMs to expand the single dynamical downscaling to a large multimodel ensemble, emulating RCM output for different GCMs.

2. Study area

This study focuses on southern half of Norway (Fig. 1), which is a region with varied geographical conditions including coastal areas with fjords and steep orography, mountain ranges, lowlands, and inland regions. The diverse geographical characteristics frame the present climate, with a maritime wet and mild climate along the west coast where temperature variations are modest, as a result of orographically forced precipitation, and a cold and dry interior with pronounced temperature variations. During the period 1971–2000, the mean 2-m temperature was 7°C near the coast in southern Norway, and -4°C in the mountains. The mean temperature

TABLE 1. Explained variance of the five leading EOF modes for various predictands and seasons, based on WRF simulations for the period 1980–2100.

Predictand	Season	EOF 1	EOF 2	EOF 3	EOF 4	EOF 5	Sum
μ	JJA	28.6	17.2	4.7	4.4	3.5	58.3
μ	SON	49.1	14.6	6.9	3.8	3	77.3
μ	DJF	60.0	12.8	6.0	2.8	2.0	83.5
μ	MAM	47.5	16.2	7.1	4.3	2.7	77.7
fw	JJA	66.5	14.5	5.1	2.0	1.1	89.2
fw	SON	59.7	24.7	3.6	1.5	1.4	90.8
fw	DJF	66.3	22.3	2.3	1.2	1.0	93.1
fw	MAM	53.2	27.8	5.3	1.6	1.5	89.5
T_{2m}	JJA	94.8	2.2	1.5	0.5	0.3	99.3
T_{2m}	SON	94.4	3.3	0.8	0.6	0.3	99.4
T_{2m}	DJF	94.2	3.6	0.8	0.6	0.2	99.4
T_{2m}	MAM	95.2	2.3	0.9	0.7	0.2	99.2

in Norway increased by 0.5°C per decade between 1976 and 2014 (Hanssen-Bauer et al. 2017).

3. Software, methods, and data

a. Empirical statistical downscaling method

The analysis was carried out within the R software environment (R Core Team 2016) with the aid of the R library “esd,” dedicated to analysis of meteorological and climate data, and empirical statistical downscaling (Benestad et al. 2007). The esd package has been used for ESD for different locations around the world, for example, Norway (Benestad et al. 2016), Poland (Mezghani et al. 2019), and Tanzania (Mtongori et al. 2016). Here the esd package was used first to aggregate the predictand (RCM output) and predictor data (GCM output) to seasonal mean values. Regression models were then calibrated as described below for each predictand variable and season [winter: December–February (DJF), spring: March–May (MAM), summer: June–August (JJA), and autumn: September–November (SON)].

Empirical orthogonal functions (EOFs) (Wilks and Wilby 1999) of the aggregated predictands y and predictors x were calculated by singular value decomposition. The EOF analysis decomposed the predictand and predictor data into a set of spatial patterns with corresponding principal components (PCs) that represent their variation in time, and eigenvalues describing their relative importance.

The decomposition of the predictand y to a finite number of orthogonal spatial patterns can be described as follows:

$$y - \bar{y} \approx \sum_{j=1}^m PCy_j \lambda_j EOFy_j. \quad (1)$$

In Eq. (1), PCy_j is the j th principal component, representing the temporal evolution of the spatial pattern described by eigenvector $EOFy_j$, and λ_j is the corresponding eigenvalue, a scaling factor describing the relative importance of PCy_j and $EOFy_j$.

A multiple regression analysis was then applied to establish a connection between the principal components of the predictand PCy_j and predictor PCx_i . A separate regression model

was fitted for each of the five leading principal components of the predictand (PCy_j for $i = 1-5$), which explain 58%–99% of the variability depending on predictand variable and season (Table 1). For simplicity and consistency, models were fitted for the same number of principal components independent of explained variance. In cases where most of the variability was in the first mode(s), the number of retained EOFs had little influence on the results because the downscaled higher-order modes add little variability.

To avoid overfitting, a stepwise forward model reduction was applied to select informative predictors among the six leading predictor modes (PCx_{ij} for $i = 1-6$). The model selection was based on the Akaike information criterion (AIC; Akaike 1974), which considers goodness of fit as well as model simplicity.

The ESD models can be summarized as follows:

$$\widehat{PCy_j} = \left(\beta_{0j} + \sum_{i=1}^{\max 6} \beta_{ij} PCx_{ij} \right), \quad (2)$$

leaving the predictands estimated by

$$\begin{aligned} \hat{y} &= \bar{y} + \sum_{j=1}^5 \widehat{PCy_j} \lambda_j EOFy_j \\ &= \bar{y} + \sum_{j=1}^5 \left(\beta_{0j} + \sum_{i=1}^{\max 6} \beta_{ij} PCx_{ij} \right) \lambda_j EOFy_j. \end{aligned} \quad (3)$$

To be able to apply the regression models to new data, a slight variation of the above described method [Eqs. (1)–(3)] was implemented in which the predictor data for the calibration period were combined with predictor data for another period and/or from another GCM simulation in a common EOF analysis (Benestad 2001). First, the new predictor data were appended to the original predictor data along its temporal axis. Then, the dominant spatiotemporal patterns of the combined field were extracted by EOF analysis [Eq. (1)]. ESD models were calibrated as described above [Eqs. (2) and (3)] using the part of the predictor principal components that represent the original predictor for the calibration period. Projections of the predictand principal components could then be obtained

by applying the same statistical model to the part of the predictor principal components representing the new data source or period. The common EOF analysis is essential to ensure that the predictor principal components used for calibration and projection represent the same spatial patterns. The projected predictand PCs can be combined with the corresponding predictand EOFs and eigenvalues from the calibration period ($\mathbf{EOF}_{y_j}$ and λ_{y_j}) to construct seasonal mean projections of the predictand field of the same spatial resolution as the original predictand data. This procedure was used 1) to obtain projections of the predictand beyond the calibration period and 2) to expand the ESD models to the full CMIP5 ensemble.

The regression models were evaluated by comparing the projections of the predictand fields with the original RCM fields for time periods excluded from calibration. This constitutes an independent validation as the predictor data used to produce these projections were not involved in the construction of the regression models.

A range of ESD models were tested by varying the predictor variable and configuration. Five GCM variables [air temperature at 2 m T_{2m} , vapor pressure at 2 m VP_{2m} , sea level pressure p_{sl} , precipitable water vapor (prw), and outgoing longwave radiation (OLR)] were tested as predictors for the three different predictands: T_{2m} , wet-day frequency f_w , and wet-day mean precipitation μ . The configuration choices that were tested included (i) calibration periods of different lengths, (ii) detrending or not detrending the principal components of the predictand and predictor prior to model calibration, and (iii) using bias-corrected or raw GCM output as predictor data. The optimal ESD model setups were then expanded to the same predictor variable from the full CMIP5 ensemble.

b. Data

A substantial part of this study involved calibrating statistical models with different setup options and evaluating them on regional climate model output in a series of “perfect model experiments.” This pseudoreality environment consisted of a GCM simulation with the Norwegian Earth System Model (NorESM1-M, ensemble member r1i1p1) given the RCP8.5 scenario for external forcing, and dynamical downscaling of the same bias-corrected GCM output covering south Norway for the period 1980–2100, as described in Pontoppidan et al. (2018). The bias correction involved adjusting the monthly mean climatology following the method of Bruyère et al. (2014), using the 1980–2010 ERA-Interim reanalysis (Dee et al. 2011) as reference data. This particular bias-correction approach was used to 1) ensure physical consistency in the final data (as opposed to bias correction in a postprocessing step) and 2) keep the climate variability from the GCM (as opposed to the pseudo-global-warming method, which assumes stationary climate variability; Schär et al. 1996; Rasmussen et al. 2011). The dynamical downscaling was done using a convection-permitting (4-km grid spacing) version of the Advanced Research version of the Weather Research and Forecasting (WRF) Model (Skamarock et al. 2008). Pontoppidan et al. (2018) found that forcing the RCM with the bias-corrected

fields increased the tilt of the North Atlantic Ocean storm track during the historical time period, producing storm tracks more consistent with those found in ERA-Interim. The spatial distribution of the winter precipitation in southern Norway was also improved in the bias-corrected dynamical downscaling. After a 20% undercatch correction was added to the observations, the dynamical downscaling based on bias-corrected GCM fields gave winter precipitation within the 90% confidence interval of the observations at 249 station locations, whereas the same held true for only 121 station locations for the dynamical downscaling based on original GCM fields (Pontoppidan et al. 2018).

The command line tools “nco” and “cdo” were used to aggregate the WRF output to seasonal mean values. A 1 mm day^{-1} threshold was used for wet-day classification to produce seasonal wet-day mean and frequency. The data were subsequently regridded to a 0.05° latitude and 0.1° longitude grid using the Python programming package xESMF (Zhuang 2018). Bilinear interpolation was used for regridding of the temperature, and a conservative interpolation scheme was used for the precipitation flux. The regridding was performed because the analysis tools used in this study are better suited to a regular longitude–latitude grid than the native rotated grid of WRF.

All CMIP5 data variables, except prw, were gathered from the KNMI Climate Explorer, where all model output had been regridded to a 2.5° by 2.5° resolution. The prw was retrieved from the ESGF servers, combining the historical runs with RCP8.5 before the fields were bilinearly remapped to a 1° by 1° grid using cdo. The regridding of the prw data was done because parts of the analysis performed in this study (the common EOF analysis in particular) that expand the single dynamical downscaling to a large multimodel ensemble requires the same spatial resolution of the predictor data used for calibration (the NorESM1-M data) and projections. An example of the spread in prw projections within the ensemble is shown in Fig. S1 in the online supplemental material. Vapor pressure VP_{2m} was calculated by combining GCM T_{2m} fields with corresponding RH fields using the Buck (1981) equation for water. Incomplete data archives in terms of GCM variable availability resulted in different ESD ensemble sizes, depending on predictor choice.

c. Metrics

The ESD model setup options were evaluated and ranked based on two metrics, the correlation score (CS) and the trend score (TS), comparing the spatiotemporal variations of the projected field with the corresponding field from the reference dataset (the dynamical downscaling output). The use of correlation as a performance metric in climate modeling is uncommon because most climate simulations do not correspond to observed variations in terms of phase and timing. However, with pseudoreality we use RCM data that do match the phase and timing of the large-scale GCM data used as predictor, for the calibration as well as the evaluation periods. Here we utilize this to extend the evaluation of the ESD results beyond the traditional range of metrics. To emphasize the dominant spatiotemporal patterns and isolate the skill of the calibration

FIG. 2. Schematic of the study, with the workflow involving input GCM data and downscaled RCM variables outlined to the left, the two calibration experiments described in the upper-center and upper-left panels, and the end product illustrated in the lower-left corner.

procedure from the effect of the EOF analysis on the predictand, the reference data were subject to EOF analysis and a reconstructed reference field based on the five leading modes of variability was used for comparison. Cross-validation strategies and application of the skill scores are described in sections 3d and 3e.

The CS was calculated as the spatial mean of temporal correlation (Pearson's r) between the predicted values ($\widehat{y}_{gc}$) and reference data (y_{gc}^*) for each grid cell (gc). The metric was calculated for one season at a time, so seasonality did not inflate the correlation scores. However, a long-term trend captured in both series would also raise the score:

$$CS = \frac{\sum_{gc=1}^N \rho(\widehat{y}_{gc}, y_{gc}^*)}{N}. \quad (4)$$

The trend score TS was defined as the fraction of grid cells for which the estimated trend was within the 95% confidence interval of the reference data trend:

$$TS = \frac{\sum_{gc=1}^N [\text{LL95} < \widehat{\text{trend}}_{gc} < \text{UL95}]}{N}, \quad (5)$$

where $\widehat{\text{trend}}_{gc}$ is the trend in the estimated field, N is the total number of grid cells, and LL95 and UL95 are the lower and upper uncertainty limits of the reference data trend given a 95% confidence interval. If the gridcell trend in the estimated field was within these limits the statement $[\text{LL95} < \widehat{\text{trend}}_{gc} < \text{UL95}]$

was 1; otherwise it was 0. Figures S7 and S8 in the online supplemental material show examples of the trend score visualized in map plots.

To reduce the computational cost, the trends were calculated using linear regression for each of the grid cells. The TS metric was chosen among others as it is bounded (between 0 and 1), unitless, and stable where reference values are near zero (as might be the case for trends in wet-day frequency). Furthermore, this metric does not result in penalization if there are no significant trends in either the reference data or the data to be evaluated and holds more information than comparing the areal mean trends.

d. Calibration experiment I: Five calibration periods to find optimal predictors

In experiment I, the five predictor variables (T_{2m} , VP_{2m} , p_{sl} , prw , and OLR) were tested as predictors for the three predictands (T_{2m} , f_w , and μ). Three of the predictor variables (T_{2m} , VP_{2m} , and p_{sl}) of the NorESM1-M model were available in two forms: (i) the original GCM fields, and (ii) as GCM fields where the mean had been bias corrected, providing a total of eight predictor fields to be tested. For each predictand and season, the predictive skill of the eight predictor fields was evaluated based on the CS and TS for different calibration and validation periods, both with and without detrending the fields prior to model calibration. A schematic illustration of the calibration experiments is included in Fig. 2.

The experiment labeled 30n used 30 years in the near present (1980–2009) for calibration and the remaining years (2010–99) for validation, while 30f used 30 years in the far future (2070–99)

for calibration and the remaining years (1980–2069) for validation. The experiment 60n used the first half (1980–2039) of the time span for calibration and the second for validation (2040–99), while 60f had the opposite periods for calibration and validation. The last experiment, c120, used the full 120-yr period (1980–2099) for both calibration and validation.

The results of experiment I was summarized by calculating a weighted mean of the metrics for the different calibration experiments (30n, 30f, 60n, 60f, and c120), where the weights were proportional to the length of the calibration time.

e. Calibration experiment II: Sensitivity to calibration range and detrending

For each predictand and season, the top two or three performing predictors from experiment I were selected for a more rigorous test of the influence of the calibration length, choice of period, and the detrending option. The calibration lengths considered were 30, 40, 50, 60, 70, 80, 90, and 100 years. For each calibration length, regression models were constructed repeatedly for all possible consecutive calibration periods within the full 120 years of available data, each time evaluating the predictive skill based on the correlation and trend scores for the full period 1980–2099. This approach provided 184 members for the 30-yr calibration length, and 44 members for the 100-yr range. Four or six experimental setups were tested for each predictand and season, amounting to 456 downscaling experiments for each experimental setup, or 29 184 experiments in total. The many experiments were performed to provide a statistical sample from which the effect of the calibration period upon the predictive skill could be evaluated. The optimal predictor variable and detrending option for each predictand was decided based on the results from the two longer calibration lengths, 90 and 100 years. However, the specific years used as calibration period in subsequent analyses were not selected based on this analysis, as this could have resulted in overfitting.

f. Expansion and clustering of the RCM ensemble

The best-performing ESD setups (predictor variable and detrending option) were selected for each predictand and season based on experiments I and II. Regression models were then fitted and applied to all available members of the CMIP5 ensemble (between 63 and 100, depending on the predictor variable) to emulate the RCM projections using the full length (1980–2099) of the predictand (WRF) and predictor (NorESM1-M) data for calibration. The common EOF analysis and stepwise multiple regression were repeated for each ensemble member to ensure that the regression model was applicable to that GCM simulation. For each predictand and season, the ensemble of projected RCMs was divided into two groups based on hierarchical clustering of the long-term trends. The clusters were used for illustrative purposes to reveal common spatial patterns of change that may have been hidden in a map of the ensemble mean (see, e.g., Fig. 6). The hierarchical clustering method (available in the R 3.3.1 base package “stats”; R Core Team 2016) was chosen because it showed good validation scores (this was tested using the R package cValid, v0.6–6; Brock et al. 2008).

4. Results

a. Impact of using bias-corrected GCM fields in ESD

The impact of utilizing bias-corrected fields in ESD was investigated by using bias-corrected GCM data (T_{2m} , VP_{2m} , and p_{sl} from the NorESM1-M model) as predictor data. Using predictor fields where the climatological monthly mean had been corrected based on reanalysis data did not result in different ESD results than those based on the original uncorrected GCM fields (see Figs. S2 and S3 in the online supplemental material).

b. Calibration experiment I: Five calibration periods to find optimal predictors

Calibration experiment I (section 3d) included five experiments with different calibration periods to find well-suited experimental setups for ESD. The three top ranking ESD experimental setups and their weighted correlation and trend scores are displayed in Fig. 3. The upper row shows the results for T_{2m} , the center for μ , and the lower for f_w . The ESD-based emulations of T_{2m} show high CS, with the best-performing setup scoring 0.83 in spring, 0.80 in summer, and 0.89 in autumn and winter. The correlation scores are generally lower for the precipitation-related variables than for temperature (particularly in summer) with the exception of the winter predictions for f_w , where the best-performing setup has a CS of 0.72. This is partly attributed to the noisiness of the precipitation field, that is, a considerable fine-scale variability that is not predictable by the ESD models. An experiment averaging the precipitation fields over regional precipitation zones showed an increase in the models' explained variance (not shown). However, the regional averaging did not increase the trend scores, and as the averaging lead to an information loss, it was considered better left as a postprocessing option. Any end user should be aware that the year-to-year variability of gridscale precipitation is only partly captured in the downscaling, particularly in summer.

The noisiness of the precipitation field can also be seen in the EOF analysis of the predictand data (Table 1). The five leading EOF modes of f_w and μ explain 89%–93% and 58%–83%, respectively, depending on season, while for T_{2m} the first leading EOF pattern accounts for 94%–95% of the variance. In summer, the five leading EOF modes of μ explain only 58% of the variability.

The trend scores for the top-performing ESD setups were quite high for all variables (0.78–0.99 for the top-performing setups). This implies that multiple regression is a suitable approach for creating ESD models replicating long term trends for both the temperature and the precipitation-related variables f_w and μ .

The best-performing experimental ESD setups for predicting T_{2m} were to use the GCM variables VP_{2m} or T_{2m} as predictor without detrending the fields prior to calibration. The same ESD setups were found to be the best for downscaling μ , but the wet-day mean could also be skillfully downscaled using the vertically integrated moisture prw, as predictor.

The best predictor for f_w was the mean sea level pressure p_{sl} in autumn and winter, whereas prw and OLR also showed fair results as predictors in spring and summer. The detrending

FIG. 3. The top three ESD experimental setups for downscaling (top) T_{2m} , (center) μ , and (bottom) f_w in each season (y axis), based on weighted averages of the metrics for calibration experiment I. Shown are the scores for the (left) first-, (center) second-, and (right) third-highest-performing setup, with the TS in the left rectangle and the CS in the right rectangle for each block. The predictor variable is denoted to the right of the two scores, in boldface italics if detrending was used.

option had very little influence on the results when p_{sl} was used as a predictor (Fig. 3).

c. Calibration experiment II: Sensitivity to calibration range and detrending

A second and more thorough investigation was pursued, including only the top-performing predictor variables for each predictand in calibration experiment I (Fig. 3 and Table 2). The ESD models were constructed using either 30, 40, 50, 60, 70, 80, 90, or 100 years for calibration, covering all available consecutive time slices within the full time period 1980–2099. To have a common reference, all models were evaluated based on their correlation and trend score for the period 1980–2099.

Figure 4 shows results of calibration experiment II for the autumn season (SON). Results for the remaining seasons are shown in Figs. S4–S6 in the online supplemental material.

The results indicate that for 30 year long calibration periods, the trend score can vary from as low as 0 up to almost 1, depending on the exact range of years (Fig. 4b). This indicates that the estimated trends depend on the calibration period and that the statistical models were not stationary. With a longer calibration period, the results became more consistent, that is, they showed less sensitivity to the specific years used for calibration. Furthermore, the skill had a tendency of being lower when detrending the predictand and predictor data prior to downscaling, especially for longer calibration periods. This finding varied, however, according to the predictor variable; sea level pressure p_{sl} showed the least sensitivity to the detrending option.

The best predictor was selected for each predictand and season based on the results from experiment II. Here, the longest calibration periods (90 and 100 year) were considered because they showed the best and most consistent performance. For the application of expanding the ESD results to other GCM runs, 120 years of RCM results were available for calibration, and skill scores based on long time periods were thus most relevant for the downscaling task at hand. Figure 5 shows the mean and standard deviation of the skill scores for the top predictors based on the members of the 90- and 100-yr moving-window calibration experiments in each season. The score values are shown as stacked bar plots, where a perfect score is two. A symbol above the bar indicates the highest ranking predictor for the predictand and season.

In some cases, very little separated the top predictor's score from the second best-performing predictor (Fig. 5). Both the large-scale fields of T_{2m} and VP_{2m} were highly skilled predictors of T_{2m} , while T_{2m} , VP_{2m} , and prw largely exhibited a similar skill as predictors for μ in each season. In autumn and

TABLE 2. The top-performing predictors for each predictand, chosen based on calibration experiment I.

Predictand	Predictor 1	Predictor 2	Predictor 3
T_{2m}	T_{2m}	VP_{2m}	—
μ	T_{2m}	VP_{2m}	prw
f_w	p_{sl}	prw	OLR

FIG. 4. Results of calibration experiment II for the autumn season (SON) for (a) CS and (b) TS for ESD models of the three predictands [(top row) T_{2m} , (second row) μ , (third row) f_w , within (a) and (b)] based on the top-performing predictors. Calibration periods of different lengths were considered (30, 40, 50, 60, 70, 80, 90, and 100 years), using a moving window covering all possible consecutive periods within the full time range of 1980–2099. The resulting distributions of CS and TS are displayed as points showing all values, as well as boxes indicating the 25th, 50th, and 75th percentile. The fill value of the boxplots indicates whether the ESD models were fitted on detrended data (fill color) or not (white fill).

FIG. 5. Stacked bar plots showing the mean CS (green fill color) and the mean TS (purple fill color) for the members within the 90- and 100-yr moving-window calibration experiments. Each panel show the result for one predictand and one predictor for each of the four seasons (x axis). The best-performing predictors are indicated with a marker above the bar. Error bars show the combined standard deviations among the members for the two skill scores.

winter, p_{sl} was clearly the most suited predictor for f_w , whereas in spring and summer prw showed marginally better results than p_{sl} .

d. Emulated RCM ensemble

The best ESD setups (predictors and detrending options) were chosen for each predictand and season based on calibration experiments I and II (see Table 3). These setups were then used to emulate the RCM results for the full CMIP5 ensembles.

The data availability for the RCP8.5 GCMs varied with the type of variable. For T_{2m} , 81 ensemble members were available, for the derived field VP_{2m} , 63 members were available, prw was available from 100 members, while the p_{sl} ensemble had 78 members. The results are shown in Fig. 6. End-of-century changes, that is, the annual 1980–2099 trends multiplied by the length of the period (120 years), are displayed over two columns, where each of the columns show the mean change in one of two clusters within the ensemble.

The emulated temperature ensemble showed positive trends in both clusters in all seasons (the two leftmost columns in Fig. 6). In winter and autumn, the temperature trends within the largest cluster showed higher temperature increases than the original dynamical downscaling of NorESM1-M (Fig. 1), but in these seasons, the ESD downscaled NorESM1-M was

clustered with the smaller group with a smaller mean change. For example, in autumn, the original dynamical downscaling showed an increase of 4.1°C per 120 years, while the larger (64%) and smaller (36%) ESD clusters indicated an average change of 6.1° and 3.6°C per 120 years, respectively. In spring and summer, the ESD downscaling of NorESM1-M was clustered with the majority of the ensemble members, which in summer had a similar average change to the original dynamical downscaling.

With regard to the wet-day mean precipitation μ , the average areal mean change was positive for both clusters in all seasons but spring (the two middle columns in Fig. 6). In summer, the larger ESD ensemble cluster showed higher increases than the smaller cluster, which included the ESD

TABLE 3. The top-performing predictors for each predictand and season based on the mean value of the sum of the two skill scores (CS and TS) for the members within the 90- and 100-yr moving-window calibration experiments.

Predictand	Autumn	Winter	Spring	Summer
T_{2m}	T_{2m}	T_{2m}	VP_{2m}	VP_{2m}
μ	prw	prw	T_{2m}	prw
f_w	p_{sl}	p_{sl}	prw	prw

FIG. 6. End-of-century changes (the annual 1980–2099 trends multiplied by the length of the period, 120 years) in seasonal T_{2m} , μ , and n_w (calculated as $f_w n$, where $n \approx 90$ is the number of days in a season) based on hybrid downscaling (ESD downscaling of RCM output) for the RCP8.5 scenario. For each of the downscaled variables, two columns of results are shown, depicting the mean change in each of two hierarchically clustered groups. In the upper-left corner of each result, the percentage of the models that are in the group and the areal mean of the group mean change are given, denoted in boldface if the group contains NorESM1-M r1i1p1.

downscaled NorESM1-M. In winter, the two clusters showed trends with different spatial patterns. The largest group (57% of ensemble members including NorESM1-M) showed little change on the west coast, whereas the smaller cluster (43%) indicated substantial coastal precipitation increases. The average areal mean change for the larger cluster was the same as the areal mean change of the dynamical downscaling (1 mm day^{-1} per 120 years), whereas the smaller cluster had an average change that was almost half as strong (0.6 mm day^{-1} per 120 years). In spring, about two-thirds of the ensemble members produced increases in precipitation, while the remaining one-third showed a weak areal mean decrease (-0.1 mm day^{-1} per 120 years).

Wet-day frequency trends are displayed as trends in the number of wet day per season; $n_w = n f_w$, where n is the number of days per season ($n \approx 90$). The clustering of the wet-day frequency trends resulted in two groups with opposite signs of areal mean change in all seasons except for spring (the two rightmost columns in Fig. 6). In winter, the larger cluster (59%) of ensemble members indicated a weak, negative trend in areal mean change (-0.2 days per 120 years), while the remaining members show a mean increase of about four wet days over the period 1980–2099. In summer, the cluster with 79% of the

ensemble members including NorESM1-M showed a decrease in f_w over eastern Norway, while in spring and autumn, the clusters with the majority of the ensemble members and NorESM1-M, indicated an average increase of about 2.5 wet days from 1980 to 2099.

5. Discussion

a. Impact of GCM bias correction prior to ESD

As expected, the ESD results were not affected by whether or not the predictor fields were climatologically biased (Figs. S2 and S3 in the online supplemental material). This may, however, be specific to the hybrid downscaling method employed in this study. As described by Eqs. (1) and (3), the ESD model predicts anomalies by making use of the variance in the predictor fields. Since the bias-correction method used in Pontoppidan et al. (2018) adjusted biases in climatological means and not variance (Bruyère et al. 2014), the bias correction did not impact the ESD analysis.

Although the GCM bias correction did not provide any added value when applied directly in the ESD analysis, the RCM results based on bias-corrected GCM fields did show added value compared to dynamical downscaling based on the

original GCM fields (see [section 3b](#)). The improved RCM data served as better ESD predictand and therefore added value in the context of a GCM–RCM hybrid dynamical–statistical downscaling approach. Since there was no need to bias correct the GCM fields, the ESD models could be calibrated with non-bias-corrected GCMs and RCM results derived from bias-corrected GCMs, and then be applied to a multimodel ensemble of non-bias-corrected GCMs.

b. Optimal predictors, sensitivity to detrending, and calibration period

1) OPTIMAL PREDICTORS

In calibration experiment I, various large-scale fields from the GCM (T_{2m} , VP_{2m} , p_{sl} , prw , and OLR) were applied as predictors for the fine-scale, dynamically downscaled predictands (T_{2m} , μ , and f_w). Of these predictor variables, T_{2m} and p_{sl} have historically most often been used in applications of ESD, while VP_{2m} , prw , and OLR have not been as common as predictors.

The predictors were evaluated by their ability to reproduce the fine-scale predictands' long term trends and annual variability (correlation), evaluated for each grid cell. For the study region, both T_{2m} and VP_{2m} were suitable for ESD of T_{2m} . VP_{2m} and prw showed similar capacities as predictors of μ , providing the highest scores in all seasons except for spring, where T_{2m} was found to provide the highest scores. Previous unpublished downscaling efforts of μ (R. Benestad 2019, unpublished material) have shown good results when OLR was used as predictor. The use of OLR was motivated by the idea that higher cloud tops often are associated with more intense precipitation and that OLR is influenced by the cloud-top height. The lesser performance when using OLR from the driving GCM as a predictor of the dynamical downscaling μ can be explained by the fact that the RCMs tend to generate a different cloud and precipitation climate than the GCM, and hence give different OLR aggregated over the same region [established from the European Coordinated Regional Climate Downscaling Experiment (Euro-CORDEX) runs (not shown)]. This implies that the GCM OLR is a less suitable predictor of the RCM's μ ; a caveat likely particular for the pseudoreality context. When ESD is applied to real observations where OLR is taken from reanalyses or satellites retrievals, the OLR field should show more consistency with local-scale μ .

2) DETRENDING

The results of calibration experiment II suggested that detrending the predictors and predictands before calibration of the ESD models affected their ability to recover the observed long-term trends for the future negatively. In this study, it was known a priori that the trends over the calibration period was part of the signal that we intended to project in the future. For real-world observations, this is not so clear-cut, and the motivation behind the detrending was that more short-term spurious trends should not interfere with the model calibration. Furthermore, detrending the data prior to model calibration enables an additional test of the models based on observations, that is, that they are able to reproduce the

trends over the calibration interval when the models use the detrended data as input. In other words, although the results indicate that the calibration data should not be detrended prior to ESD, there may in some contexts be good reasons for doing so.

3) SENSITIVITY TO CALIBRATION LENGTH AND PERIOD

The impact of both calibration length and choice of calibration years were scrutinized in calibration experiment II. The purpose of the experiment was to provide information about whether the ESD models calibrated in one period were valid for another, in a changing climate, that is, to investigate the question of model stationarity. RCP8.5 is considered as an upper trajectory under a high-emissions scenario, and thus involves a highly nonstationary climate. Testing ESD stationarity under RCP8.5 is thus an "extreme" test of model stationarity.

The results ([Fig. 5](#), along with Figs. S4–S6 in the online supplemental material) showed that the sensitivities varied widely depending on the metric considered, predictor choice, detrending option, and predictand. Generally, a higher sensitivity to calibration length was found when the ability to reproduce the long term trend was evaluated rather than the year-to-year variability. A minimum calibration length may be indicated by a stabilization of the metrics and low spread among the calibration length experiment members. For most ESD setups, a minimum of 60 years or even longer was required to get a stable model. In real-world applications, the availability of observation based data limits the calibration length.

A caveat of experiment II is that the skill scores were calculated on all available data (1980–2099), meaning that the proportion of independent validation data decreased with the length of the calibration period. To address this issue, another version of experiment II was conducted where for each calibration period, the CS was calculated using all data not used for calibration, that is, only independent data. The results (not shown) were similar to the results described above and the analysis supports the conclusions regarding long calibration periods and detrending.

The value of a long calibration period has been known before and is, for example, the reason why many ESD studies have used the older NCEP–NCAR R2 or ERA-40 reanalysis rather than a newer ones (e.g., ERA-Interim) for calibration, as the former cover a longer period in time. The hybrid approach implemented here overcomes the problem of lacking observational data limiting ESD calibration length. It also provides calibration data not only from historical time, but from a future projection, which likely contains the effects of local feedback effects that might not yet have taken place in historic time. A caveat of this approach, however, is that it relies on the ability of the dynamical downscaling to accurately capture these nonlinear changes.

Future studies, including more than one constitutive centennial dynamical downscaled GCM run, might shed light on the sensitivity to RCM settings within the hybrid approach framework. Furthermore, future work should include exploring if any added value is seen when an additional layer of complexity is added, namely constructing empirical statistical

downscaling models, by not only combining the GCM and RCM fields, as in the hybrid approach, but also including information from the traditional pairing of reanalysis data and high-resolution observation fields. Additional use of observational data in the hybrid approach could also be used to bias correct the RCM data before training the statistical models.

c. The emulated RCM ensemble

The emulated RCM ensemble, based on results from 63 to 100 GCM members, showed a considerable spread, even when clustered into just two groups (Fig. 6). In many cases the majority of the emulated ensemble showed centennial trends deviating from those from the original downscaled RCM. For seasonal wet-day frequency f_w , the two clusters of the pseudo-RCM ensemble showed diverging trends in large areas of the study domain in three of four seasons. The emulated RCM ensemble has been deposited at the data-storage facility Zenodo (<https://doi.org/10.5281/zenodo.3552397>), and the Norwegian Meteorological Institute's (MET Norway) thredds server (<http://thredds.met.no/thredds/catalog/metusers/helenebe/catalog.html>), providing easy access for further use. It should be noted that, while the long-term trend was well produced by all downscaled variables, the statistical model showed less skill in reproducing the interannual variability on a gridcell basis for the precipitation variables (μ and f_w). Since aggregation to precipitation zones showed higher skill cores for reproducing interannual variability, spatial averaging to a lower resolution, municipality scale, or catchment areas may provide more robust estimates when interannual variability is concerned. Given that different users will prefer different aggregation zones, and that the long-term trend is fairly well reproduced at gridcell resolution, any spatial aggregation is left to the end user.

The variability within the pseudo-RCM results provides valuable information for climate adaptation and communication. In this study, seasonal statistics mean temperature, wet-day mean precipitation, and wet-day frequency were downscaled. From the two latter variables additional precipitation statistics may be derived (Benestad et al. 2019). The results of the hybrid downscaling method will likely provide a robust and reliable contribution to the downscaling and impact research communities. The hybrid approach is an efficient downscaling method useful for expanding dynamical downscaling results, which may become a more important task as dynamical downscaling moves to a convection resolving resolution with an increasing computational cost.

The application of the ESD method to the entire CMIP5 ensemble depends on the common EOF analysis, which ensures that the same spatial patterns are found in the GCM data used for calibration (the NorESM1-M data) and the ensemble member to which the regression is applied. An inspection of the common EOFs shows similar statistical characteristics of the parts of the common EOF PCs representing NorESM1-M and the CMIP5 members.

6. Conclusions

The availability of 120 years of GCM and RCM data under the RCP8.5 scenario provided the opportunity to test the ESD

method and model setup and identify the best-performing predictor choice and configuration options during a period where the climate is nonstationary. The stationarity of the statistical model was tested by varying the calibration and validation periods. Furthermore, the sensitivity to using climatologically bias-corrected GCM fields was been investigated. The extensive experiments undertaken allowed us to state the following findings, at least valid within the current implementation of empirical statistical downscaling in a “perfect model experiment” setting:

- 1) Using GCM fields, where the mean has been climatologically bias corrected, did not improve the ESD estimates.
- 2) Detrending the data prior to calibration reduced the skill of the ESD (Fig. 4).
- 3) A 30-yr calibration period is often not sufficient for skillful model calibration (Fig. 4). In general, ESD skill increased with calibration length, particularly the trend scores. In the pseudoreality setting that we have investigated here, even with a calibration period of 50 years it is possible to get a highly unskilled statistical model (trend score 0).

The results point to that any extra effort to extract as long time series as possible for calibration will be rewarded by producing more robust statistical models. The hybrid approach has value as it can make way for long calibration periods, and also the possibility to calibrate the statistical models on the time period for which they will be produced for. The latter point is of importance, as the local tipping points of feedback effects influencing the variables to be downscaled, may not yet have occurred in historical time. Most machine learning methods will likely show a similar sensitivity to choice and length of calibration time, and some, for example, the random forest method implemented in Shortridge et al. (2016), cannot predict values beyond those seen in the calibration period.

The rigorous model selection included in this study has likely resulted in a higher veracity of the ESD of the GCM ensemble. Future work might entail work on developing ESD models combining long time series of observation–reanalysis and RCM–GCM data to further test and optimize the models. Additional dynamically downscaled results for the study region would also enable more stringent validation of the emulated RCM ensemble and facilitate an investigation of the sensitivity to the choice of calibration and validation datasets.

Acknowledgments. This study is part of RCubed (Grant 255397), an innovative project founded by the Research Council of Norway, in which statistical and dynamical downscaling communities came together with social scientists, meteorologists, and hydrologists to provide more resilient, robust, and reliable climate projections for the future. We thank those involved in an early part of the project, in particular Stefan Sobolowski and Mathew Alexander Stiller-Reeve, and we thank project partners at NVE and NORCE for valuable discussions. We thank Martin King (NORCE) for his contribution to the RCM simulations and RCN's program for supercomputing (NOTUR/NORSTORE) Projects NN9280K, NN9486K, and NS9001K for computational resources to the

RCM simulations. We also thank Andreas Dobler (MET Norway) for making the RCM data available and our local IT support for assisting us in using our local postprocessing infrastructure.

REFERENCES

- Akaike, H., 1974: A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723, <https://doi.org/10.1109/TAC.1974.1100705>.
- Bartók, B., and Coauthors, 2017: Projected changes in surface solar radiation in CMIP5 global climate models and in EURO-CORDEX regional climate models for Europe. *Climate Dyn.*, **49**, 2665–2683, <https://doi.org/10.1007/s00382-016-3471-2>.
- Bellprat, O., S. Kotlarski, D. Lüthi, and C. Schär, 2012: Objective calibration of regional climate models. *J. Geophys. Res.*, **117**, D23115, <https://doi.org/10.1029/2012JD018262>.
- Benestad, R. E., 2001: A comparison between two empirical downscaling strategies. *Int. J. Climatol.*, **21**, 1645–1668, <https://doi.org/10.1002/joc.703>.
- , A. Mezghani, and K. M. Parding, 2007: ESD: Climate analysis and empirical-statistical downscaling (ESD) package for monthly and daily data. Tech. Rep., 272 pp., <https://github.com/metno/esd>.
- , K. M. Parding, K. Isaksen, and A. Mezghani, 2016: Climate change and projections for the Barents region: What is expected to change and what will stay the same? *Environ. Res. Lett.*, **11**, 054017, <https://doi.org/10.1088/1748-9326/11/5/054017>.
- , —, H. B. Erlandsen, and A. Mezghani, 2019: A simple equation to study changes in rainfall statistics. *Environ. Res. Lett.*, **14**, 084017, <https://doi.org/10.1088/1748-9326/ab2bb2>.
- Berg, N., A. Hall, F. Sun, S. Capps, D. Walton, B. Langenbrunner, and D. Neelin, 2015: Twenty-first-century precipitation changes over the Los Angeles region. *J. Climate*, **28**, 401–421, <https://doi.org/10.1175/JCLI-D-14-00316.1>.
- Brock, G., V. Pihur, S. Datta, and S. Datta, 2008: clValid: An R package for cluster validation. *J. Stat. Software*, **25**, 1–22, <https://doi.org/10.18637/jss.v025.i04>.
- Bruyère, C. L., J. M. Done, G. J. Holland, and S. Fredrick, 2014: Bias corrections of global models for regional climate simulations of high-impact weather. *Climate Dyn.*, **43**, 1847–1856, <https://doi.org/10.1007/s00382-013-2011-6>.
- Buck, A. L., 1981: New equations for computing vapor pressure and enhancement factor. *J. Appl. Meteor.*, **20**, 1527–1532, [https://doi.org/10.1175/1520-0450\(1981\)020<1527:NEFCVP>2.0.CO;2](https://doi.org/10.1175/1520-0450(1981)020<1527:NEFCVP>2.0.CO;2).
- Charles, S. P., B. C. Bates, P. H. Whetton, and J. P. Hughes, 1999: Validation of downscaling models for changed climate conditions: Case study of southwestern Australia. *Climate Res.*, **12**, 1–14, <https://doi.org/10.3354/cr012001>.
- Cornes, R. C., G. van der Schrier, E. J. M. van den Besselaar, and P. D. Jones, 2018: An ensemble version of the E-OBS temperature and precipitation data sets. *J. Geophys. Res. Atmos.*, **123**, 9391–9409, <https://doi.org/10.1029/2017JD028200>.
- Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, <https://doi.org/10.1002/qj.828>.
- Deser, C., A. Phillips, V. Bourdette, and H. Teng, 2012: Uncertainty in climate change projections: The role of internal variability. *Climate Dyn.*, **38**, 527–546, <https://doi.org/10.1007/s00382-010-0977-x>.
- Dixon, K. W., J. R. Lanzante, M. J. Nath, K. Hayhoe, A. Stoner, A. Radhakrishnan, V. Balaji, and C. F. Gaitán, 2016: Evaluating the stationarity assumption in statistically downscaled climate projections: Is past performance an indicator of future results? *Climatic Change*, **135**, 395–408, <https://doi.org/10.1007/s10584-016-1598-0>.
- Enriquez-Alonso, A., J. Calbó, A. Sanchez-Lorenzo, and E. Tan, 2017: Discrepancies in the climatology and trends of cloud cover in global and regional climate models for the Mediterranean region. *J. Geophys. Res. Atmos.*, **122**, 11 664–11 677, <https://doi.org/10.1002/2017JD027147>.
- Fernández, J., and Coauthors, 2019: Consistency of climate change projections from multiple global and regional model inter-comparison projects. *Climate Dyn.*, **52**, 1139–1156, <https://doi.org/10.1007/s00382-018-4181-8>.
- Hanssen-Bauer, I., and Coauthors, 2017: Climate in Norway 2100—A knowledge base for climate adaptation. NCCS Rep. 1/2017, 48 pp., <https://www.miljodirektoratet.no/globalassets/publikasjoner/M741/M741.pdf>.
- IPCC, 2013: *Climate Change 2013: The Physical Science Basis*. Cambridge University Press, 1535 pp., <https://doi.org/10.1017/CBO9781107415324>.
- Kotlarski, S., and Coauthors, 2017: Observational uncertainty and regional climate model evaluation: A pan-European perspective. *Int. J. Climatol.*, **39**, 3730–3749, <https://doi.org/10.1002/JOC.5249>.
- Lanzante, J. R., K. W. Dixon, M. J. Nath, C. E. Whitlock, and D. Adams-Smith, 2018: Some pitfalls in statistical downscaling of future climate. *Bull. Amer. Meteor. Soc.*, **99**, 791–803, <https://doi.org/10.1175/BAMS-D-17-0046.1>.
- Li, G., X. Zhang, F. Zwiers, and Q. H. Wen, 2012: Quantification of uncertainty in high-resolution temperature scenarios for North America. *J. Climate*, **25**, 3373–3389, <https://doi.org/10.1175/JCLI-D-11-00217.1>.
- Mezghani, A., A. Dobler, R. Benestad, J. E. Haugen, K. M. Parding, M. Piniewski, and Z. W. Kundzewicz, 2019: Subsampling impact on the climate change signal over Poland based on simulations from statistical and dynamical downscaling. *J. Appl. Meteor. Climatol.*, **58**, 1061–1078, <https://doi.org/10.1175/JAMC-D-18-0179.1>.
- Mtongori, H. I., F. Stordal, and R. E. Benestad, 2016: Evaluation of empirical statistical downscaling models' skill in predicting Tanzanian rainfall and their application in providing future downscaled scenarios. *J. Climate*, **29**, 3231–3252, <https://doi.org/10.1175/JCLI-D-15-0061.1>.
- Parding, K. M., R. Benestad, A. Mezghani, and H. B. Erlandsen, 2019: Statistical projection of the North Atlantic storm tracks. *J. Appl. Meteor. Climatol.*, **58**, 1509–1522, <https://doi.org/10.1175/JAMC-D-17-0348.1>.
- Pontoppidan, M., E. W. Kolstad, S. Sobolowski, and M. P. King, 2018: Improving the reliability and added value of dynamical downscaling via correction of large-scale errors: A Norwegian perspective. *J. Geophys. Res. Atmos.*, **123**, 11 875–11 888, <https://doi.org/10.1029/2018JD028372>.
- , —, S. P. Sobolowski, A. Sorteberg, C. Liu, and R. Rasmussen, 2019: Large-scale regional model biases in the extratropical North Atlantic storm track and impacts on downstream precipitation. *Quart. J. Roy. Meteor. Soc.*, **145**, 2718–2732, <https://doi.org/10.1002/qj.3588>.
- Rasmussen, R., and Coauthors, 2011: High-resolution coupled climate runoff simulations of seasonal snowfall over Colorado: A process study of current and warmer climate. *J. Climate*, **24**, 3015–3048, <https://doi.org/10.1175/2010JCLI3985.1>.
- R Core Team, 2016: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, <https://www.R-project.org/>.

- Schär, C., C. Frei, D. Lüthi, and H. C. Davies, 1996: Surrogate climate-change scenarios for regional climate models. *Geophys. Res. Lett.*, **23**, 669–672, <https://doi.org/10.1029/96GL00265>.
- Schmith, T., 2008: Stationarity of regression relationships: Application to empirical downscaling. *J. Climate*, **21**, 4529–4537, <https://doi.org/10.1175/2008JCLI1910.1>.
- Shortridge, J. E., S. D. Guikema, and B. F. Zaitchik, 2016: Machine learning methods for empirical streamflow simulation: A comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrol. Earth Syst. Sci.*, **20**, 2611–2628, <https://doi.org/10.5194/hess-20-2611-2016>.
- Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., <https://doi.org/10.5065/D68S4MVH>.
- Sun, F., A. Hall, M. Schwartz, D. B. Walton, and N. Berg, 2015: Twenty-first-century snowfall and snowpack changes over the Southern California mountains. *J. Climate*, **29**, 91–110, <https://doi.org/10.1175/JCLI-D-15-0199.1>.
- Takayabu, I., H. Kanamaru, K. Dairaku, R. Benestad, H. von Storch, and J. H. Christensen, 2015: Reconsidering the quality and utility of downscaling. *J. Meteor. Soc. Japan*, **94A**, 31–45, <https://doi.org/10.2151/jmsj.2015-042>.
- 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**, L18701, <https://doi.org/10.1029/2007GL030295>.
- Walton, D. B., F. Sun, A. Hall, and S. Capps, 2015: A hybrid dynamical–statistical downscaling technique. Part I: Development and validation of the technique. *J. Climate*, **28**, 4597–4617, <https://doi.org/10.1175/JCLI-D-14-00196.1>.
- , A. Hall, N. Berg, M. Schwartz, and F. Sun, 2017: Incorporating snow albedo feedback into downscaled temperature and snow cover projections for California’s Sierra Nevada. *J. Climate*, **30**, 1417–1438, <https://doi.org/10.1175/JCLI-D-16-0168.1>.
- Wilby, R. L., 1997: Non-stationarity in daily precipitation series: Implications for GCM down-scaling using atmospheric circulation indices. *Int. J. Climatol.*, **17**, 439–454, [https://doi.org/10.1002/\(SICI\)1097-0088\(19970330\)17:4<439::AID-JOC145>3.0.CO;2-U](https://doi.org/10.1002/(SICI)1097-0088(19970330)17:4<439::AID-JOC145>3.0.CO;2-U).
- Wilks, D., and R. Wilby, 1999: The weather generation game: A review of stochastic weather models. *Prog. Phys. Geogr.*, **23**, 329–357, <https://doi.org/10.1177/030913339902300302>.
- Zhuang, J., 2018: xESMF: Universal Regridder for Geospatial Data. <https://github.com/JiaweiZhuang/xESMF>.