Climatic Change (2016) 135:395–408
DOI 10.1007/s10584-016-1598-0

Evaluating the stationarity assumption in statistically downscaled climate projections: is past performance an indicator of future results?

Keith W. Dixon[1] · **John R. Lanzante**[1] · **Mary Jo Nath**[1] · **Katharine Hayhoe**[2] · **Anne Stoner**[2] · **Aparna Radhakrishnan**[3] · **V. Balaji**[4] · **Carlos F. Gaitán**[5]

Received: 2 July 2015 / Accepted: 3 January 2016 / Published online: 22 January 2016

Abstract Empirical statistical downscaling (ESD) methods seek to refine global climate model (GCM) outputs via processes that glean information from a combination of observations and GCM simulations. They aim to create value-added climate projections by reducing biases and adding finer spatial detail. Analysis techniques, such as cross-validation, allow assessments of how well ESD methods meet these goals during observational periods. However, the extent to which an ESD method's skill might differ when applied to future climate projections cannot be assessed readily in the same manner. Here we present a "perfect model" experimental design that quantifies aspects of ESD method performance for both historical and late 21st century time periods. The experimental design tests a key stationarity assumption inherent to ESD methods – namely, that ESD performance when applied to future projections is similar to that during the observational training period. Case study results employing a single ESD method (an Asynchronous Regional Regression Model variant) and climate variable (daily maximum temperature) demonstrate that violations of the stationarity assumption can vary geographically, seasonally, and with the amount of projected climate change. For the ESD method tested, the greatest challenges in downscaling daily maximum temperature projections are revealed to occur along coasts, in summer, and under conditions of

Electronic supplementary material The online version of this article (doi:10.1007/s10584-016-1598-0) contains supplementary material, which is available to authorized users.

✉ Keith W. Dixon
Keith.Dixon@noaa.gov

[1] NOAA Geophysical Fluid Dynamics Laboratory, 201 Forrestal Road, Princeton, NJ 08540, USA

[2] Climate Science Center, Texas Tech University, Lubbock, TX 79409, USA

[3] Engility, Chantilly, VA 20151, USA

[4] Princeton University, Princeton, NJ 08544, USA

[5] University of Oklahoma, Norman, OK 73072, USA

greater projected warming. We conclude with a discussion of the potential use and expansion of the perfect model experimental design, both to inform the development of improved ESD methods and to provide guidance on the use of ESD products in climate impacts analyses and decision-support applications.

1 Introduction

Global climate models (GCMs) play an important role in advancing the scientific understanding of large-scale climate variations and trends, including those observed over the past century. When driven by plausible changes in radiative forcing agents, GCMs generate a range of future climate projections. Yet, many who wish to incorporate climate projections into adaptation planning and decision making applications find GCM-generated data inadequate for direct use (Gleick 1986; Wigley et al. 1990; Snover et al. 2013). This is especially true when focusing on regional or local-scale issues (Giorgi and Mearns 1991; Huth et al. 2000). Among the commonly cited shortcomings associated with GCM data products are the lack of fine-scale spatial resolution and biases in the GCM-simulated 20th century climate relative to observations (Benestad et al. 2008). To partly address these shortcomings, empirical statistical downscaling (ESD) techniques may be applied to refine GCM-generated climate projections (Wilby and Wigley 1997).

There are a broad range of ESD methods of varying levels of complexity (Benestad et al. 2008; Hewitson et al. 2014). At their core, they generally use as input GCM simulations and observation-based datasets to determine statistical relationships that in turn are applied to transform GCM outputs into downscaled products. In practice, ESD outputs typically are viewed as value-added products – deemed to be more credible and suitable for downstream applications than the raw GCM results from which they are derived. However, assumptions that may limit the suitability of downscaled projections for some applications often are not conveyed to or appreciated by end users (Hall 2014, Hewitson et al. 2014). For past time periods, ESD performance characteristics can be determined by comparing observational datasets with ESD-generated products representing the same time period (e.g., via cross-validation (Bishop 2006; Wilks 2011)). However, lacking future observations, assessing the credibility of ESD-generated projections for climate conditions several decades in the future poses significant challenges (Barsugli et al. 2013).

One critical assumption implicit to all ESD methods is that of *statistical stationarity*, which presumes the statistical relationships between GCM output and observed climate data utilized by ESD techniques to produce downscaled projections remain constant over time (Wilby and Wigley 1997;). Though this assumption is sometimes acknowledged, studies attempting to test its validity have been limited (Frías et al. 2006; Vrac et al. 2007; Hertig and Jacobeit 2008; Maraun 2012; Gutierrez et al. 2013; Hawkins et al. 2013; Hertig and Jacobeit 2013; Gaitan and Cannon 2013; Teutschbein and Seibert 2013; Gaitan et al. 2014). In section 2, we describe an evaluation framework that seeks to isolate and quantitatively assess aspects of this *stationarity assumption*, which presumes an ESD technique's performance during the recent past to be indicative of its performance when applied to future climate projections. We refer to the evaluation framework employed in this study as a *perfect model* (*PM*) experimental design.

In section 3, we illustrate the PM experimental design's utility by presenting results focusing on how well the stationarity assumption holds for one ESD method's downscaling of late 21st century daily maximum temperatures. Section 4 contains a summary of our case study results and discussion of how adoption of the PM evaluation framework can aid efforts to assess the credibility of statistically downscaled projections commonly used to support adaptation planning and decision making.

2 A 'perfect model' experimental design

Our experimental design protocol may be considered an example of a *perfect model* approach. Note that the name is not meant to imply that the model itself is perfectly free of errors. Rather, it is a name given to a general experimental design approach in which, for analysis purposes, model-generated data serves as a substitute for observations or truth. A key element of our PM approach is the substitution of high resolution climate model results for the observational datasets typically used in ESD applications. In the PM framework, use of model simulations as proxies for observations enables one to evaluate quantitatively ESD performance in a consistent manner for both historical and future time periods. Our implementation aims to isolate uncertainties associated with the stationarity assumption that are inherent to the generation of future climate projections via any ESD technique.

Here we examine time series of daily maximum temperatures for a region centered on the conterminous 48 United States. However, the experimental design can be applied to a wide variety of climate variables, sampling frequencies, geographic locations, and ESD methods. Also, we envision that modifications of the basic experimental design can highlight different aspects of ESD method performance.

Figure 1 depicts, in schematic form, the types of datasets and methodological steps used to generate statistically downscaled climate projections in typical applications (Fig. 1a) and in the PM framework (Fig. 1b). The datasets and methods used in this study are described in the following synopsis, which contrasts a more typical implementation of ESD methods to that used in the PM experimental design.

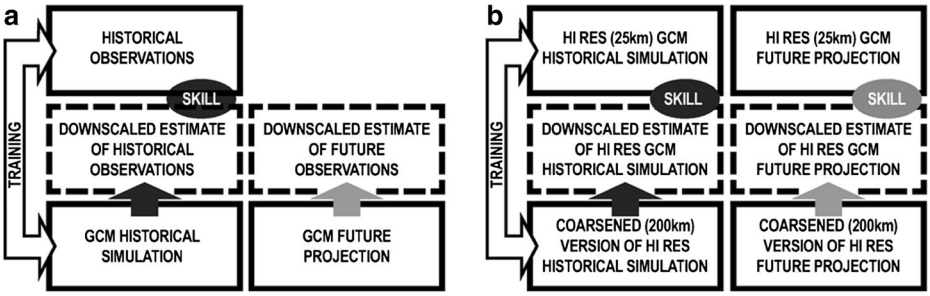

Fig. 1 Schematic representations of datasets and procedures associated with **a** a real-world implementation of ESD techniques and **b** an implementation of this study's perfect model experimental design. In both **a** and **b**, transform functions are computed in the ESD training step using datasets from a historical period as input (*white arrows*). Next, the transform functions are used to generate downscaled results for the historical (*black arrow*) and future periods (*gray arrow*). Measures of the ESD method's skill in the historical period (*black oval*) can be computed directly (e.g., via cross-validation). In the PM case **b**, use of high resolution GCM output as a proxy for "truth" allows computation of an ESD method's skill for the future projections (*gray oval*). Comparing skill scores for the historical and future periods provides information on how well the stationarity assumption holds

2.1 Outline of a typical ESD application

The three solid boxes in Fig. 1a (labeled "Historical Observations", "GCM Historical Simulation", and "GCM Future Projection") represent the input data sets available when generating ESD products. Regardless of the particular ESD technique, information gleaned via statistical methods from some combination of these data sets is used to define transfer functions – statistical relationships that associate the large-scale GCM predictors and local-scale observational predictands.

In the Fig. 1a schematic, the process begins with a pair of data types for the historical period - one representing observations for the locations and climate variable of interest and the other being GCM outputs for the same period. During the ESD training step, statistical methods are used to derive transfer functions relating the modeled and observed datasets. For the historical era, cross-validation tests may be performed to produce a set of downscaled estimates (Fig. 1a dashed box labeled "Downscaled Estimate of Historical Observations"). In cross-validation, the historical data is successively partitioned into "training" and "validation" segments. For each partition, the transfer functions derived from the training sample are applied to the independent validation sample to create downscaled estimates. Comparing the observational dataset with the downscaled historical estimates allows one to assess the ESD method's performance or skill. The skill is a measure of how well the ESD method accounts for GCM biases and shortcomings in capturing finer scale details for the places and times for which there are observations.

Having completed the ESD training step using historical era datasets, the next step is to apply the transfer function to the GCM's future climate projections. Inserting the output of future GCM projections (used as predictors) into the downscaling equations (transfer functions) established during the historical period-based training step yields downscaled versions of the future GCM projections (Fig. 1a dashed box labeled "Downscaled Estimate of Future Observations") – a process that presumes a GCM's future projections contain biases similar to those found in the GCM's historical simulation. Statistically downscaled climate projections produced in this manner are used in many climate impacts studies. However, lacking observations of the future, one cannot readily assess how well the ESD transfer functions derived from historical data perform when applied to future GCM projections; hence the question of whether ESD skill is degraded when applied to future projections typically is left unaddressed.

2.2 Perfect model data sets and methods

The data sets used in the PM experimental design (Fig. 1b) differ from those used in a conventional real-world ESD application (Fig. 1a), but the general workflow is the same. For the perfect model ESD training step, output from a relatively high resolution general circulation model serves as the predictand (Fig. 1b "Hi Res (25 km) GCM Historical Simulation"). Years 1979–2008 serve as the historical period in our case study, though other climatological periods can be used. Predictors are derived from the same high resolution GCM, but are degraded (i.e., processed to yield spatially smoothed or coarsened fields). In our case study, the high resolution general circulation model is the GFDL-HiRAM-C360 model (hereafter C360) and its output is stored on a ~ 25 km grid. Two historical C360 ensemble members, each 30 years in length, are used. (See Online Resource 1 for additional information on the C360 model experiments.)

To create the predictors labeled "Coarsened (200 km) Version of Hi Res" in Fig. 1b, C360 model output is first interpolated to a grid with ~200 km spacing using a first-order conservative scheme. There is one grid point on the 200 km grid (Fig. 2b) for every 64 grid points on the C360 grid (Fig. 2a). In a second interpolation step, the 200 km gridded fields are interpolated back to the high resolution C360 grid (Fig. 2c), to facilitate subsequent processing. For multiple C360 model simulations of both the historical and future periods, the two-step interpolation process (hereafter referred to as "coarsening") is applied to yield spatially smoothed predictor fields.

As shown in Fig. 2, much of the finer scale detail present in the original C360 predictand fields is lost during the interpolation processing. Though area-means of predictands and coarsened predictors averaged over sufficiently large regions are nearly identical, the interpolation steps can produce biases at smaller spatial scales, with biases being more pronounced in areas of strong horizontal gradients, such as along strong fronts, areas of steep elevations changes, and some coastal locations (Fig. 2d and Online Resource 2).

In the PM experimental design depicted in Fig. 1b, the ESD training step compares a predictand drawn directly from a C360 historical period simulation to coarsened predictors derived from the same experiment. Following a cross-validation approach, downscaled estimates (Fig. 1b dashed box labeled "Downscaled Estimate of Hi Res GCM Historical Simulation") are created using an independent sample of coarsened predictors as input to the transfer functions. This allows one to quantify, for the historical period, the effectiveness of an ESD method's transfer functions at recovering finer scale details of the C360 model data that the coarsening process obscured.

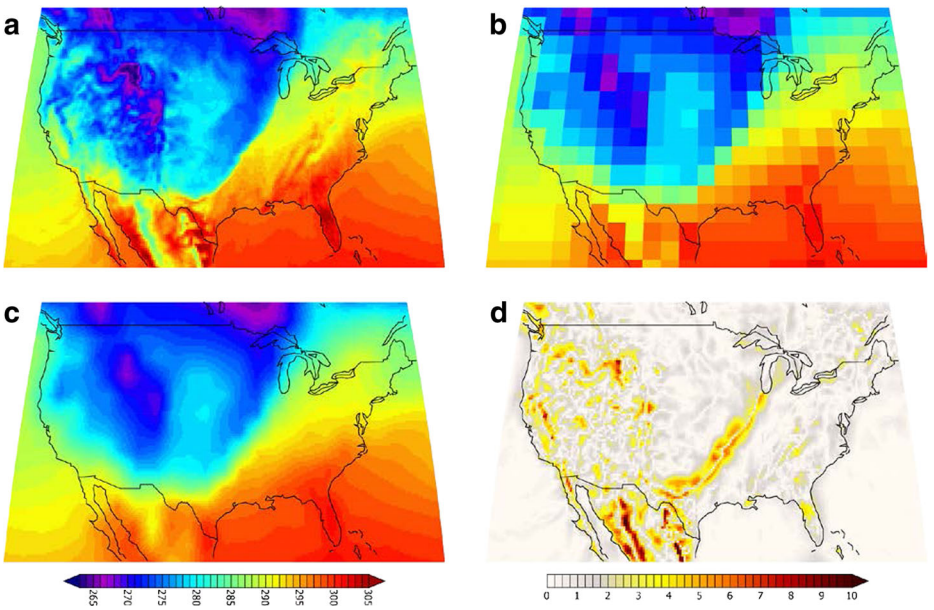

Fig. 2 Maximum surface air temperatures (tasmax, [K]) for a single autumn day characterized by a strong cold front in the central United States. **a** As simulated in a GFDL-HiRAM-C360 experiment and stored on a ~ 25 km grid. **b** After interpolation to a grid with ~200 km spatial resolution. **c** After interpolating from **b** back to the original ~25 km grid. **d** Absolute difference |**c** minus **a**|

Next, downscaled future projections (Fig. 1b dashed box labeled "Downscaled Estimate of Hi Res GCM Future Projection") are created by using coarsened future predictors (degraded versions of the C360 high resolution data) as input to the previously derived ESD transfer functions. For this case study, the future period is 2086–2095, though other future periods could be used. A key aspect of the PM framework is the availability of the "Hi Res (25 km) GCM Future Projection" data sets (upper right box of Fig. 1b). Differencing the downscaled future projections with the C360 model's original output allows one to quantify, for the future period, how effective an ESD method's transfer functions are at recovering information contained in the original C360 model data sets that was lost in the coarsening process used to develop the future predictor fields.

The stationarity assumption is presumed to hold if downscaling skill metrics associated with a future projection are statistically indistinguishable from cross-validation results computed from the companion historical period. Likewise, if the downscaled results for the future period exhibit significantly larger errors (less skill) than for the historical period, that provides a measure of how the stationarity assumption does not hold (Vrac et al. 2007). In this way, the PM approach outlined here provides a framework to test the extent to which the statistical relationships (transfer functions) determined during ESD training steps remain valid when used to generate downscaled future climate projections.

2.3 Perfect model experiments using the ARRM downscaling method

Here we demonstrate the utility of the PM experimental design by examining how one particular ESD method performs within the PM framework for one climate variable. The ESD method examined is a variant of the asynchronous regional regression model (ARRM version 1) of Stoner et al. (2013) and the variable of interest is daily maximum near-surface air temperature (tasmax). The ARRM is one of a widely-used class of ESD methods that operate on the distributional characteristics of data samples. Its workflow is consistent with that depicted in Fig. 1b. The transfer relation created by the ARRM method is a piecewise linear regression function between quantiles of high resolution C360 data (the predictand) and the coarsened version of the historical data (the predictor). In this setting, the points to which the fitting is performed are such that predictor/predictand pairs represent values that share the same relative position in their respective cumulative probability distribution functions.

The ARRM variant used here differs from that used in most ARRM applications in that an option to identify gross outliers in the downscaled output and replace them with "missing values" was disabled for this study. Thus, for some small fraction of locations and times, we retained downscaled values that other ARRM implementations would flag as suspicious and eliminate from the final product. Our choice ensures that downscaled values are available for evaluation at all locations and times; however, this leads to some larger downscaling errors being included in our analyses than would be the case had this option not been disabled.

We conducted four sets of experiments – two for the historical period (cross-validated and non-cross-validated), and two late 21st century cases. A notable difference between the two future era ensembles is that one (ensemble "C") on average exhibits about 2 K more warming than does the other (ensemble "E") (see Online Resource 1).

A summary of the four experiment types examined in this study follows.

- Hist 60lo0 case: The 1979–2008 historical era case conducted without cross-validation (i.e., with no independent sample; hence, *60* years *l*eave *o*ut *0*). During the ARRM training

step, datasets from two 30-year long C360 ensemble members are pooled together so the training step operates on 60 years of data.

- Hist 2-fold case: The 1979–2008 historical era case conducted with 2-fold cross-validation. Having two 30 year historical ensemble members, the ARRM training first uses high resolution C360 and spatially coarsened values from one 30 year ensemble member (the dependent sample) to develop downscaling transfer functions. Next, coarsened predictors from the other historical ensemble member (the independent sample) are input to the transfer functions to generate downscaled values. The process is repeated, swapping the independent and dependent time series, yielding a total of 60 years of downscaled output.
- Future E case: The 2086–2095 future era case based on C360 experiments run under the RCP8.5 radiative forcing scenario and forced using sea surface temperature anomalies from the GFDL-ESM2M model. As was done for the Hist 60lo0 case, the ARRM training step pools all 60 historical era years to determine the downscaling transfer equations. The transfer equations are used with coarsened tasmax predictors from each of three 10-year ensemble members (the E ensemble) to generate a total of 30 years of downscaled future projections.
- Future C case: The 2086–2095 future era case based on C360 experiments run under scenario RCP8.5 and forced using sea surface temperature anomalies from the GFDL-CM3 model. The method used for the Future E case is followed, substituting three 10-year C ensemble experiments as input to the training and downscaling steps.

See Online Resource 1 for additional information about the C360 climate model experiments from which the PM predictors and predictands were derived.

3 Results

Results presented here are drawn from the four sets of PM experiments described in section 2.3, in which a version of the ARRM method is used to downscale daily maximum temperatures (variable tasmax). These case study results illustrate that the extent to which the stationarity assumption holds can vary geographically, by season, and by the amount of climate change exhibited in the future GCM simulation relative to the historical period. These analyses do not comprise an exhaustive examination of the ARRM method performance nor should they necessarily be considered representative of ESD methods in general.

3.1 The magnitude of the perfect model downscaling challenge

In the perfect model framework, the ESD method attempts to recover details in the high resolution C360 data that were lost during the coarsening process used to create the predictors. The degradation associated with the coarsening process can be quantified by computing, grid point by grid point and day by day, absolute differences between high resolution C360 data values (the "target" or "truth") and the corresponding coarsened predictor values ($|$*coarsened predictor – high resolution target*$|$; hereafter $|\Delta_{pt}|$). Averaged over all days and all land points in the domain, the $|\Delta_{pt}|$ statistic is very similar for the historical period (1.51 K), the Future E ensemble (1.50 K) and the Future C ensemble (1.47 K). These quantities are depicted as black bars in Fig. 3a. The similarity of these three values indicates the coarsening process presents an ESD method with approximately the same challenge in each of the three cases, if it is to

perfectly recover the high resolution "truth". Online Resources 2 and 3 provide more information on the quantitative nature of the challenge the PM datasets pose to an ESD method, including discussion of seasonal and geographic variations.

3.2 Area mean and geographic distributions of historical and future downscaling errors

The Mean Absolute Error (MAE) metric (average of |*downscaled output – high resolution target*|) is used as a measure of downscaling skill. Color-filled bars in Fig. 3a indicate the area-averaged downscaling MAE metric computed over land points and averaged over all days. A value of zero indicates that the ESD technique perfectly recovers all information in the high resolution C360 data lost during the coarsening process used to create the predictors. Conversely, if the downscaling MAE metric is not less than the corresponding $|\Delta_{pt}|$ value (black bar) then the downscaling process could be said to not have added value.

The downscaling MAE computed for the Future C ensemble is greater than that computed for the Future E ensemble, which in turn is greater than the downscaling MAE for the historical period cases (Fig. 3a). This indicates that the ARRM downscaling adds value (downscaling MAE values are less than corresponding $|\Delta_{pt}|$ averages) but does not perform

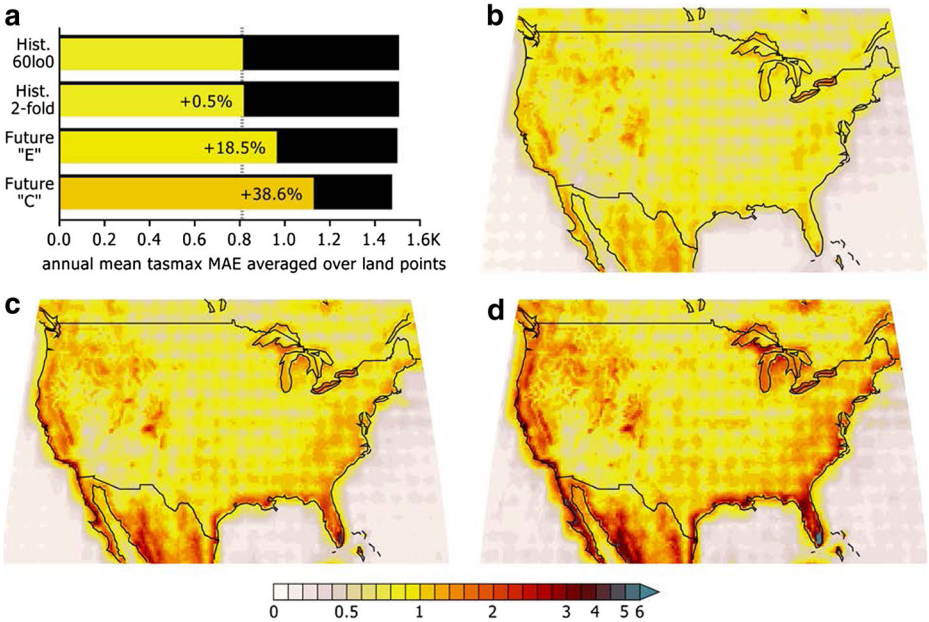

Fig. 3 Annual-average mean absolute errors (MAE, [K]) calculated for the ARRM method's downscaling of variable tasmax in the PM framework. The color bar applies to all panels (note irregular intervals). **a** Black bars represent the land area-averaged mean absolute temperature difference before downscaling $|\Delta_{pt}|$, based on differences between the daily coarsened tasmax values (predictors) and corresponding high resolution C360 tasmax values (targets). Color-filled bars represent post-downscaling land area-averaged MAEs for four experiments (based on differences between downscaled daily values and corresponding targets). Percentages listed indicate downscaling MAE increase relative to the Hist 60lo0 experiment. **b** Annual mean downscaling MAE computed over the 60 years of the historical era Hist 2-fold case. **c** Annual mean downscaling MAE computed over the 30 years of the Future E case. **d** Annual mean downscaling MAE computed over the 30 years of the Future C case

as well when downscaling tasmax under conditions of large climate change as it does during the historical era. This is evidence that the stationarity assumption was violated when generating late 21st century downscaled projections. In the C360 model simulations, future E(C) ensemble tasmax values over land average about 5.0(7.2)K warmer than during the historical period and downscaling MAE values are about 18.5(38.6)% greater, suggesting that ESD performance degrades nonlinearly, with errors increasing more quickly as temperatures increase and become more dissimilar to the historical training period**.** Robust rank order tests (Siegel and Castellan, 1988) performed on the area-averaged daily MAE time series indicate that both the 18.5 % and 38.6 % increases in the area-averaged MAE values depicted in Fig. 3a are statistically significant at the 5 % level.

Geographic distributions of the downscaling MAE, averaged over all days, are shown in Fig. 3 for three experiment types. In the future projections, larger MAE values tend to be found along coasts and some steep mountainous areas. For both late 21st century cases, the ARRM method has the most difficulty (largest MAE) downscaling tasmax at a grid point near Miami, Florida, yielding an annually-averaged MAE of 7.3 K in the warmer C ensemble.

3.3 Examining stationarity using MAE ratios

Examination of ratios computed as the downscaling MAE of a future case divided by the downscaling MAE of the cross-validated historical case provides information on the extent to which the stationarity assumption holds in the PM framework (Fig. 4, MAE values are averaged over time and region of interest before calculating ratios.) An MAE ratio of 1.0 indicates no degradation in downscaling performance under changing climate conditions, according to this metric. Ratios greater than 1.0 occur when the future ensemble's MAE exceeds that of the historical ensemble, with a ratio of 2.0 representing a doubling of the future ensemble's MAE relative to the historical case.

In Fig. 4a, seasonal variations in late 21st century downscaling skill are apparent in MAE ratios based on daily averages computed over all land points and time-filtered via a 31-day running mean. The greatest proportional increase in land area-averaged absolute errors appears in summer months. For both the future E and C ensembles, the largest time-filtered summer-time MAE ratios are more than double their respective annual averages (dashed lines in Fig. 4a). Comparatively little ESD performance degradation (ratios near 1.0) occurs during winter. Across all months, land-averaged MAEs of the warmer future C ensemble are approximately twice that of the future E ensemble, indicating that the non-linear degradation in ARRM downscaling performance as the projected climate warms is not limited to summer.

Marked geographic variations exist in MAE ratios computed from annual-mean absolute downscaling errors (Fig. 4c,d). MAE ratios less than 1.1 signal that, for this ARRM method variant, the stationarity assumption holds fairly well over large portions of the central USA, even when applied to late 21st century conditions having daily maximum temperatures that are much higher than those in the historical era used for training purposes (see Online Resource 1 for C360 projected future warming patterns). Larger annual mean MAE ratios (greater than 1.5 and 2.0 in the future E and C ensembles, respectively) tend to be found near the coast, illustrating that not only are large absolute downscaling errors characteristic of coastal regions for late 21st century tasmax projections (Fig. 3d), but that under conditions of marked warming downscaling errors proportionally grow more quickly in coastal than interior locations.

Although comparison of the future E and C ensemble results reveals that the stationarity assumption breaks down more quickly as the projected warming increases, the amount of

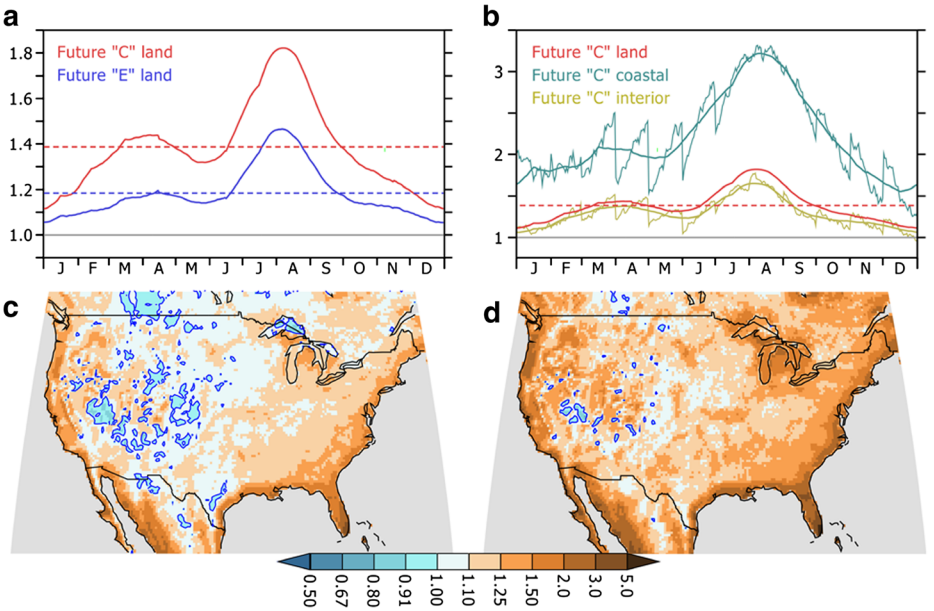

Fig. 4 Variations in the ARRM method's performance in the PM framework expressed as the ratio of MAE values. Ratios are calculated by dividing future ensemble MAEs for variable tasmax by MAEs of the historical 2-fold cross-validation case. A ratio of 1.0 indicates no degradation of downscaling performance for future conditions (i.e., the stationarity assumption holds). **a** Seasonally-varying ratios computed by dividing the area-averaged MAEs computed over all land grid points in the future C ensemble (*red*) and future E ensemble (*blue*) by the historical ensemble's land area-averaged MAEs. Time means are computed for each climatological calendar day and a 31-day running mean applied as a temporal filter. Annual means are shown as dashed lines. **b** Same as **a**, but for geographic subsets of the future C ensemble. The cyan curves show area-averaged values computed over coastal locations only (those that are three grid points or less distant from the nearest ocean point). Orange curves depict area-averaged MAE ratios for interior (non-coastal) land points. Unsmoothed curves show MAE ratios for each of the 365 climatological days. For reference, the smooth red curve and red dashed line, representing MAE ratios averaged over all land points, are duplicated from panel **a**. The sawtooth pattern evident in the unsmoothed curves is described in section 3.4. **c** Annual mean MAE ratios calculated at each land grid point for the future E ensemble. **d** Same as **c**, but for the future C ensemble. The color bar applies to panels **c** and **d**

warming alone is not a reliable indicator of how well the stationarity assumption will hold. For example, for the C ensemble the projected tasmax warming in Iowa is ~3 K greater than that projected for southeastern Florida, yet Iowa's average MAE ratio (<1.15) is much less than that of the coastal portion of Florida's Broward and Miami-Dade counties (>5.0). Similarly, locations with large time-mean differences between the original C360 tasmax data and the coarsened predictors at the corresponding grid point ($|\Delta_{pt}|$) are not reliable indicators of how well the stationarity assumption holds. The largest values of this quantity are found in mountainous regions of the western United States, not coastal locations. (See Online Resource 2 for related information.)

3.4 Considerations of when future predictors lie beyond the ARRM training predictor range

This case study's PM-based results suggest that downscaling errors tend to grow when future predictor values input to the ARRM method's downscaling transfer equations lie outside the

range of historical predictor values used in the training step. This is consistent with results presented in Figs. 3 and 4 showing (a) the area and time-averaged downscaling MAE is greater for the future C ensemble than the future E ensemble (Fig. 3), (b) the stationarity assumption holds least well during the warmest summer months, as indicated by the time-filtered MAE ratio metric (Fig. 4a), (c) downscaling errors in future projections are greater during the warmer portions of spring and autumn months, leading to a "sawtooth" pattern in the daily MAE ratio time series (Fig. 4b), and (d) the stationarity assumption tends to hold less well along coastlines than further inland (Fig. 3c,d and Fig. 4b).

Aspects of the four downscaling error behaviors mentioned above may be influenced by data preparation techniques employed in the ARRM downscaling method, as described in Section 2.3 of Stoner et al. (2013). For the training step, historical period target and predictor data sets are divided into 12 groups, each centered on a calendar month and extended two weeks on each side. For each calendar month, transfer equations are derived for each grid point and applied to downscale tasmax values for dates in the central month. The ±2 week time window expansion approximately doubles the sample size used in the training process, promoting more stable statistics. Additionally, the ±2 week extensions lead to a wider range of values being sampled for training purposes. All else being equal, downscaling accuracy tends to decrease when predictors used to generate downscaled output lie outside the bounds of those used during ARRM method training step. So, the expansion of the training time window to encompass tasmax values warmer than those of the central month yields transfer equations that can be better suited for application to future projections. However, the greater the projected warming is, the greater the probability that a future tasmax predictor will be warmer than the warmest historical predictor used in the training step. Thus, it is not surprising that the ARRM variant tested here yields larger downscaling errors for the late 21st century C ensemble than for the E ensemble, and that downscaling errors were smallest for the historical period.

The seasonal cycle of downscaling MAE ratios seen in Fig. 4a,b can be related to the prevalence of future tasmax predictor values that exceed the maximum value of historical predictors used in the training step for a particular month. For the warmest calendar month in the historical climatology, expanding the training time window ±2 weeks is unlikely to add many warmer tasmax values to the sample. Thus, for the warmest summer month, a larger fraction of predictors in the late 21st century projections can be expected to exceed the maximum historical tasmax predictor used in the training step than is the case for other months, contributing to the stationarity assumption holding least well during the peak of summer.

The sawtooth pattern evident in daily climatologies of downscaling MAE ratios computed for future projections (Fig. 4b) is a byproduct of the way the ARRM method performs its training one month at a time and can be linked to the same "beyond the training bounds" factor cited above. The future projections' MAE ratios are notably larger during the warm end of transitions months (e.g., late April, early November) than during cooler parts of those months. This intra-month pattern is consistent with the expectation that the frequency and extent to which future tasmax predictors exceed the warmest historical predictor used in the ESD training step grows as one considers projections having larger climate change signals.

That in the examples presented here, the stationarity assumption is most readily violated in late 21st projections along coastal regions is associated in part with relatively low variances in the temporal distributions of historical and future coarsened predictors at those locations. In the PM framework, spatial interpolations that are part of the coarsening process used to create the predictors effectively spread maritime influences inland, often causing the variance of tasmax predictor values at a coastal land points to be smaller than the co-located predictand's variance.

A similar effect can exist in GCM-based predictors used in typical ESD applications, as described in Online Resource 3. If one assumes a future warming signal appears as a uniform shift of the tasmax distribution, less projected warming is required for coarsened predictors to exceed the bounds of historical predictors used in the ARRM training step for a given month – conditions shown to be challenging for the ESD method examined here.

The reduced predictor variance explanation offered above (and illustrated further in Online Resource 3) is not the sole factor that challenges ESD applications in some coastal areas. Sea breeze effects, differential heating of land and ocean, and regional climate dynamics are factors that can vary along coastlines as climate changes (Hall 2014) in ways that can be difficult for an ESD method to capture in training steps that compare historical era GCM output with observations.

4 Discussion and future work

This paper presents a perfect model experimental design for the quantitative evaluation of the stationarity assumption in statistically downscaled climate projections. The PM evaluation framework allows one to address the critical question of whether the performance of an ESD method for past climate conditions is indicative of the skill it exhibits when applied to future climate projections. Illustrative results demonstrate how the PM framework allows strengths and weaknesses of an ESD method to be identified that could not be readily determined from analyses based solely on historical observations. For example, results indicate that for late 21st century projections of daily maximum temperatures in the conterminous United States, the ESD method tested in this case study tends to exhibit markedly larger downscaling errors along coastal regions and in the warmest summer months. In contrast, the stationarity assumption tends to hold better in many interior locations and during winter months.

The results presented here are not exhaustive nor intended to be taken as representative of all ESD methods or all climate variables. This paper's primary goals are to describe the PM experimental design framework and to illustrate its value by examining one climate variable and one ESD method. Informed by these results, potential modifications to the ARRM downscaling method have been tested, yielding some performance improvement with respect to the stationarity assumption. Note that the ARRM version used in this study differs somewhat from that documented in Stoner et al. (2013) in that a post-processing step that operates on suspected outliers was omitted here. Tests (not shown) indicate our case study findings are not sensitive to that omission.

Future plans include adapting the experimental design to confront multiple ESD methods with a wider range of stationarity assumption challenges and analyzing the results with an expanded set of metrics. This may involve, but is not limited to, incorporating data sets from other high resolution GCMs, examining different time periods and radiative forcing scenarios, altering GCM-based input files in systematic ways to focus on particular features, and the use of synthetic time series. Ongoing work (not shown) utilizing the PM framework has yielded preliminary results suggesting that notable performance differences exist across ESD methods and across different climate variables and indices. To facilitate tests of stationarity assumption performance in other ESD methods, the PM input datafiles used in this case study are available to the research community (see http://www.gfdl.noaa.gov/esd).

We anticipate that the potential wider application and adaptation of this PM evaluation framework could enhance the informed exchange of data and knowledge between the physical

climate science and climate impacts research communities, while simultaneously promoting the development of improved ESD techniques. Results of this type can provide valuable information regarding the level of confidence one should attribute to ESD-generated climate projections commonly used in impacts analyses and as the basis for decision-support and planning purposes.

Acknowledgments Our work benefitted from discussions with several National Climate Predictions and Projections Platform workshop participants, Claudia Tebaldi, and Isaac Held. We thank Vaishali Naik, Gabriel Vecchi, and two anonymous reviewers for their constructive comments. The effort to develop and implement this perfect model evaluation framework received funding from the U.S. Geologic Survey's South Central Climate Science Center (cooperative agreements G12AC20512 and G13AC00387) and the National Oceanic and Atmospheric Administration's (NOAA) Climate Program Office. V. Balaji is supported by the Cooperative Institute for Climate Science, Princeton University, under NOAA award NA08OAR4320752.

References

Barsugli JJ et al. (2013) The practitioner's dilemma: how to assess the credibility of downscaled climate projections. Eos Trans Amer Geophys Union 94:424–425. doi:10.1002/2013EO460005

Benestad RE, Hanssen-Bauer I, Chen D (2008) Empirical-statistical downscaling. World Scientific Publishing Company, New Jersey

Bishop CM (2006) Pattern recognition and machine learning. Springer, New York

Frías MD, Zorita E, Fernández J, Rodríguez-Puebla C (2006) Testing statistical downscaling methods in simulated climates. Geophys Res Lett 33:L19807. doi:10.1029/2006GL027453

Gaitan CF, Cannon AJ (2013) Validation of historical and future statistically downscaled pseudo-observed surface wind speeds in terms of annual climate indices and daily variability. Renew Energy 51:489–496. doi:10.1016/J.Renene.2012.10.001

Gaitan CF, Hsieh WW, Cannon AJ (2014) Comparison of statistically downscaled precipitation in terms of future climate indices and daily variability for southern Ontario and Quebec, Canada. Clim Dyn 43(12). doi:10.1007/s00382-014-2098-4

Giorgi F, Mearns LO (1991) Approaches to the simulation of regional climate change: a review. Rev Geophys 29:191. doi:10.1029/90RG02636

Gleick PH (1986) Methods for evaluating the regional hydrologic impacts of global climatic changes. J Hydrol 88:97–116. doi:10.1016/0022-1694(86)90199-X

Gutierrez JM, San-Martin D, Brands S, Manzanas R, Herrera S (2013) Reassessing statistical downscaling techniques for their robust application under climate change conditions. J Clim 26(1):171–188. doi:10.1175/JCLI-D-11-00687.1

Hall A (2014) Projecting regional change. Science 346:1461–1462. doi:10.1126/science.aaa0629

Hawkins E, Osborne TM, Ho CK, Challinor AJ (2013) Calibration and bias correction of climate projections for crop modelling: an idealised case study over Europe. Agric For Meteorol 170:19–31. doi:10.1016/j.agrformet.2012.04.007

Hertig E, Jacobeit J (2008) Assessments of Mediterranean precipitation changes for the 21st century using statistical downscaling techniques. Int J Climatol 28(8):1025–1045. doi:10.1002/joc.1597

Hertig E, Jacobeit J (2013) A novel approach to statistical downscaling considering nonstationarities: application to daily precipitation in the Mediterranean area. J Geophys Res 118(2):520–533. doi:10.1002/jgrd.50112

Hewitson BC, Daron J, Crane RG, et al. (2014) Interrogating empirical-statistical downscaling. Clim Chang 122:539–554. doi:10.1007/s10584-013-1021-z

Huth R, Kyselý J, Pokorná L (2000) A GCM simulation of heat waves, dry spells, and their relationships to circulation. Clim Chang 46:29–60. doi:10.1023/A:1005633925903

Maraun D (2012) Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums. Geophys Res Let 39:L06706. doi:10.1029/2012GL051210

Siegel S, Castellan NJ (1988) Nonparametric statistics for the behavioral sciences. McGraw-Hill, New York

Snover AK, Mantua NJ, Littell JS, et al. (2013) Choosing and using climate-change scenarios for ecological-impact assessments and conservation decisions. Conserv Biol 27:1147–1157. doi:10.1111/cobi.12163

Stoner AMK, Hayhoe K, Yang X, Wuebbles DJ (2013) An asynchronous regional regression model for statistical downscaling of daily climate variables. Int J Climatol 33:2473–2494. doi:10.1002/joc.3603

Teutschbein C, Seibert J (2013) Is bias correction of regional climate model (RCM) simulations possible for non-stationary conditions? Hydrol Earth Syst Sci 17:5061–5077. doi:10.5194/hess-17-5061-2013

Vrac M, Stein ML, Hayhoe K, Liang X-Z (2007) A general method for validating statistical downscaling methods under future climate change. Geophys Res Lett 34:L18701. doi:10.1029/2007GL030295

Wigley TML, Jones PD, Briffa KR, Smith G (1990) Obtaining sub-grid-scale information from coarse-resolution general circulation model output. J Geophys Res: Atmos 95:1943. doi:10.1029/JD095iD02p01943

Wilby RL, Wigley TML (1997) Downscaling general circulation model output: a review of methods and limitations. Prog Phys Geogr 21:530–548. doi:10.1177/030913339702100403

Wilks DS (2011) Statistical methods in the atmospheric sciences, 3rd edn. Academic Press, Waltham, MA