An objective statistical downscaling technique for emulating WRF

George Shu Heng Pau[1], Daniel B. Walton[*2], Samir Touzani[3]

[1]Lawrence Berkeley National Laboratory

Climate and Ecosystem Sciences Division

1 Cyclotron Road

Berkeley, CA 94720.

[2]University of California, Los Angeles

Institute of the Environment and Sustainability

La Kretz Hall, Suite 300

Los Angeles, CA 90095-1496.

[3]Lawrence Berkeley National Laboratory

Building Technology and Urban Systems Division

1 Cyclotron Road

Berkeley, CA 94720.

*Corresponding author email address: waltond@ucla.edu

Abstract

Accurate downscaling of global climate models (GCMs) is needed to quantify the local impacts of climate change. Dynamical downscaling with a regional climate model has been shown to capture important physical processes at fine scales, but it is too computationally expensive to be used to downscale a large ensemble of GCMs or multiple time periods and scenarios. Hybrid dynamical-statistical downscaling saves time by using a statistical method to mimic the output of dynamical downscaling. Previous applications of hybrid downscaling used a subjective statistical method to fit the region of interest. It is preferable to use an objective, automated statistical technique that is easily portable to any region. Here, Proper Orthogonal Decomposition Mapping (PODM) is presented as a potential candidate. As a case study, PODM is used to mimic output from the Weather Research and Forecasting (WRF) model used to project climate changes over California's Sierra Nevada mountain range. The results show that PODM robustly predicts WRF temperatures from coarse GCM output, with similar errors across different GCM cases. PODM predictions have 15% lower error than the original hybrid model of Walton et al. (2016). More importantly, PODM can be implemented using automated procedures with limited manual tuning, allowing it to be deployed rapidly. PODM is also shown to compare favorably to state-of-the-art machine-learning algorithms in the context of hybrid downscaling. The use of an objective statistical technique like PODM has the potential to streamline the application of hybrid downscaling for other regions.

1 Introduction

Predictions from global climate models (GCMs) are commonly used to study the impacts of climate change on various aspects of human activities (Gosling et al. 2011). However, existing climate models do not necessarily have the required resolution to accurately model relevant fine-scale features, such as complex topography in mountain ranges and urban heat island effects in cities (McCarthy et al. 2010). While ongoing efforts attempt to resolve these features directly within global climate models, downscaling procedures are practical approaches that allow us to obtain climate change predictions at the desired resolution of a particular impact study.

Downscaling techniques have been widely used to downscale climatic variables, typically precipitation and temperature, from global to regional scales; these techniques have been well-documented, e.g., in Benestad et al. (2008), Fowler et al. (2007), Gutmann et al. (2014), Maraun et al. (2010), and Wilby et al. (1998). There are two main approaches: dynamical downscaling and the statistical downscaling. Dynamical downscaling simulates the complex physical processes that underlie the local climate response using a regional climate model (RCM) forced at its boundaries by reanalysis or GCM output. Statistical downscaling uses a statistical model to map coarse GCM output to station observations or a gridded dataset. A wide variety of statistical downscaling methods are available. For example, Bias Correction with Spatial Disaggregation (Wood et al. 2004), has been particularly successful for downscaling of precipitation suitable as input to regional hydrological models. More complex regression models (Hanssen-Bauer et al.

2003; von Storch et al. 1993) have been used to directly model the relationship between the predictors (e.g., sea level pressure) and climatic variables of interest. A similar approach called pattern scaling (Tebaldi and Arblaster 2014) has been used within integrated assessment models.

More recently, hybrid dynamical-statistical downscaling techniques have been developed (referred to here as "hybrid downscaling"). Hybrid downscaling uses a statistical model to extend the results of dynamical downscaling to multiple GCMs. Under this approach, dynamical downscaling is applied to a small subset of GCMs. Then, a statistical model is trained to mimic the dynamically downscaled results, and is applied to the remaining GCMs. This saves time when downscaling a large ensemble of GCMs, as applying a statistical models is typically much faster than performing dynamical downscaling. Hybrid downscaling may be valuable in situations where there are important features of the climate change pattern that can only be captured through dynamical downscaling (Berg et al. 2015; Sun et al. 2015a; Sun et al. 2015b).

The statistical models used in previous hybrid downscaling by Walton et al. (2015) and Walton et al. (2016) require the user to manually investigate the dynamically downscaled data and parameterize salient processes affecting the climate change signal in the region of interest. An open question is whether an objective statistical method could be used instead to minimize manual tuning and streamline the process. Furthermore, previous statistical models were designed to downscale only changes in climatology, but it would be desirable to be able to downscale time series as well. Here we investigate whether

Proper Orthogonal Decomposition Mapping (PODM; Pau et al., 2014) — a method that was successfully used to downscale hydrological and biogeochemical quantities in Pau et al. (2016) — could be used in hybrid downscaling. The aim of this paper is to determine whether PODM can accurately and robustly emulate dynamically downscaled temperatures when fed coarse GCM output. We also systematically investigate the effects of using different predictors and predictands.

2 Problem Setup

In Walton et al. (2016), the authors downscaled GCM climate change projections for California's Sierra Nevada mountain range. To capture the effects of complex topography and snow albedo feedback (SAF) on the warming in the Sierra Nevada, high-resolution (3 km) simulations were performed with the Weather Research and Forecasting model (WRF; Skamarock et al., 2008). Following a hybrid approach, five GCMs were dynamically downscaled with WRF. Then, the WRF climate change patterns were used to train a statistical model, called StatWRF, that was used to produce WRF-like climate change patterns for an entire ensemble of 35 GCMs. Here we follow a similar procedure, but using PODM and machine learning techniques instead of StatWRF to extend the WRF results.

As we use the Walton et al. (2016) WRF simulations, it is necessary to briefly describe them. WRF version 3.5 (Skamarock et al. 2008) is used in a configuration with three one-way nested domains of 27, 9, and 3 km resolution, going from the outermost to innermost domain (Figure 1). WRF was coupled to the community Noah land surface model with multi-parameterization options (Noah-MP) (Niu et al. 2011). First, a historical simulation was performed, with WRF forced by North American Regional Reanalysis (NARR; (Mesinger et al. 2006) spanning the period 1991-2000. (NARR data was provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their website at https://www.esrl.noaa.gov/psd/data/gridded/data.narr.html.) Next, five "future" simulations were performed, each representing how the 1991-2000 period would have transpired if the mean climate were altered by changes between the 2081-2100 and 1981-2000 periods in a different CMIP5 GCM (Taylor et al. 2012) run under the RCP 8.5 scenario (Riahi et al. 2011). The five GCMs used are CNRM-CM5, GFDL-CM3, INMCM4, IPSL-CM5A-LR, and MPI-ESM-LR (see acronym details at http://www.ametsoc.org/PubsAcronymList). Each future simulation is forced with boundary conditions created by adding the difference in GCM monthly climatology (2081–2100 minus 1981–2000) to the 1991–2000 NARR data. This process was applied to temperature, humidity, zonal and meridional winds, and geopotential height. Readers should refer to Walton et al. (2016) for a full description of the WRF model and the dynamical downscaling step. WRF temperature data used in this study is available from http://research.atmos.ucla.edu/csrl/pub.html.

In the statistical downscaling step, we employ PODM and machine learning techniques to determine the high-resolution WRF monthly 2m air temperature (T) for the innermost domain (D3) from low-resolution GCM output. Our predictand is $T_{\mathrm{WRF,fut}}$, the sequences of monthly T values in the "future" simulations. This is a more difficult task than using the change in climatology, $\Delta\overline{T}_{\mathrm{WRF}}=\overline{T}_{\mathrm{WRF,fut}}-\overline{T}_{\mathrm{WRF,hist}}$, as the predictand, as is considered in Walton et al. (2016) because the statistical model must be able to explain inter-annual variability, not just mean changes. In above, $\overline{T}_{\mathrm{WRF,fut}}$ and $\overline{T}_{\mathrm{WRF,hist}}$ are the 10-year average of the monthly temperature in the 2091-2100 and 1991-2000 periods, respectively.

Four predictors are considered as input: the monthly NARR 2m air temperature from 1991-2000 (T_{NARR}), the monthly NARR surface temperature from the same period ($T_{\mathrm{S,NARR}}$), the difference in GCM monthly 2m air temperature climatology between the 2081-2100 and 1981-2000 periods ($\Delta\overline{T}_{\mathrm{GCM}}$), and the difference in GCM monthly surface temperature climatology between the 2081-2100 and 1981-2000 periods ($\Delta\overline{T}_{\mathrm{S,GCM}}$). [Note that the differences in monthly climatology have length 12, while the time series for 1991-2000 has length 120 (10 years $\times$ 12 months/year). So, the sequence of differences in monthly climatology are repeated 10 times when serving as a predictor.] Since the resolutions between NARR and GCM are different, $\Delta\overline{T}_{\mathrm{GCM}}$ and $\Delta\overline{T}_{\mathrm{S,GCM}}$ are interpolated to the resolution of NARR data.

We also consider an alternative way of preparing the predictors that is more similar to the way the future WRF boundary conditions are constructed. Under this alternate preparation, T_{NARR} and $\Delta\overline{T}_{\mathrm{GCM}}$ are combined into a single predictor

$$T_{\mathrm{BC,fut}}=T_{\mathrm{NARR}}+\Delta\overline{T}_{\mathrm{GCM}}\,. \tag{1}$$

Similarly, we define $T_{\mathrm{S,BC,fut}} = T_{\mathrm{S,NARR}} + \Delta\overline{T}_{\mathrm{S,GCM}}$. As part of the statistical model setup, we determine which set of predictors (and which way of preparing them) minimizes the error. The goal is to see whether surface temperature should be included along with 2m air temperature T, and whether it is advantageous to use the combined predictors such as $T_{\mathrm{BC,fut}}$ that mimics the future WRF boundary conditions.

In addition, we consider direct and indirect approaches to obtaining $T_{\mathrm{WRF,fut}}$ from the above predictors. The direct approach is to use $T_{\mathrm{WRF,fut}}$ as the predictand. The indirect approach is to use the temperature change $\Delta T_{\mathrm{WRF}} = T_{\mathrm{WRF,fut}} - T_{\mathrm{WRF,hist}}$ as the predictand and then add the result to the historical sequence of temperatures $T_{\mathrm{WRF,hist}}$ to determine $T_{\mathrm{WRF,fut}}$. The indirect approach could be useful since it matches the way the boundary conditions are constructed, i.e. by adding the climate change signal to historical sequence.

To compare our results with StatWRF (which was designed to downscale climatological changes), we also test PODM's skill in mimicking $\Delta\overline{T}_{\mathrm{WRF}}$. We are interested to see if PODM can improve on the accuracy of StatWRF while still capturing the temperature sequences.

3 Methods

3.1 Proper orthogonal decomposition mapping (PODM)

We give a summary of the PODM method, as formulated in Pau et al. (2016). This method was first proposed by Robinson et al. (2006) and is derived from the Gappy proper orthogonal decomposition (POD) method (Everson and Sirovich 1995). We first consider a single multivariate predictor $\mathbf{p}$ (e.g. $T_{\mathrm{BC,fut}}$ over the region D1) and a single multivariate predictand $\mathbf{f}$ (e.g. ΔT_{WRF} over the region D3). The training dataset consists of N snapshots of $\mathbf{p}$ and $\mathbf{f}$, taken monthly over the 10-year simulation period using different GCM outputs. For example, N is 600 if output from the the five future simulations is used (each simulation yields 120 monthly snapshots). Given N corresponding sets of these $\mathbf{p}$ and $\mathbf{f}$ snapshots, we determine a set of POD bases that are found through a singular value decomposition of the following data matrix:

$$\mathbf{W}^{\mathrm{PODM}} = \begin{bmatrix} \mathbf{p}_1 - \overline{\mathbf{p}} & \cdots & \mathbf{p}_N - \overline{\mathbf{p}} \\ \mathbf{f}_1 - \overline{\mathbf{f}} & \cdots & \mathbf{f}_N - \overline{\mathbf{f}} \end{bmatrix} \tag{2}$$

where $\overline{\mathbf{p}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{p}_i$ and $\overline{\mathbf{f}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{f}_i$. We determine M right singular vectors, $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_M\}$ corresponding to the M largest singular values for the above data matrix. The POD bases are then given by $\zeta_i = \mathbf{W}^{\mathrm{PODM}}\mathbf{v}_i, i = 1, \ldots, M$, and represent dominant modes of variability in the snapshots within the data matrix $\mathbf{W}^{\mathrm{PODM}}$. By decomposing ζ_i into

$$\zeta_i = \begin{bmatrix} \zeta_i^{\mathbf{p}} \\ \zeta_i^{\mathbf{f}} \end{bmatrix} \tag{3}$$

where $\zeta_i^{\mathbf{p}}$ and $\zeta_i^{\mathbf{f}}$ are components of the POD basis vector associated with the predictor and the predictand. A linear approximation of **f** in the vector space spanned by ζ_i is then given by

$$\mathbf{f} \approx \mathbf{f}_{\mathrm{PODM}} = \overline{\mathbf{f}} + \sum_{i=1}^{M} \gamma_i \zeta_i^{\mathbf{f}} \ . \tag{4}$$

The PODM method determines $\gamma = \{\gamma_1, \ldots, \gamma_M\}$ that solves the following the least square problem:

$$\gamma = \arg\min_{\alpha} \| \mathbf{p} - \overline{\mathbf{p}} - \sum_{i=1}^{M} \alpha_i \zeta_i^{\mathbf{p}} \|_2 \tag{5}$$

where $\| \cdot \|_2$ is the discrete L^2 measure. We expect the PODM model to be good if $\zeta_i^{\mathbf{p}}$ is strongly correlated to $\zeta_i^{\mathbf{f}}$. We can augment the above procedure to consider multiple predictors by stacking multiple predictors into a single vector, i.e.

$$\mathbf{p} = \begin{bmatrix} \mathbf{p}^1 \\ \mathbf{p}^2 \\ \vdots \\ \mathbf{p}^n \end{bmatrix}, \tag{6}$$

assuming we have n different predictors. In this study, $n = 4$.

A key parameter of PODM model is the number of POD bases (M) used in the approximation. Determining an appropriate M is a balance of accuracy and stability. If M is too large, the PODM approximation becomes unstable due to over-fitting (Everson and Sirovich 1995; Pau et al. 2014). If M is too small, then PODM will not capture all modes of variability and accuracy will be diminished. To determine an appropriate *M*, we will utilize a leave-one-out cross validation (LOOCV) procedure. There are six total cases that we can use for training: results from the five future simulations and the historical case, which represents the zero change scenario (i.e. $\Delta\overline{T}_{\text{GCM}} = 0$ and $\Delta\overline{T}_{\text{S,GCM}} = 0$). Under LOOCV, we train different a PODM model using the historical case and four out of the five GCM cases, with each of the five GCM cases taking a turn being left out. Each PODM model is then used to predict the T_{WRF} for the case not included in the training dataset. For each case, we determine the mean absolute error (MAE) for a range of *M*. An "optimal" *M*, M_{opt}, is given by one that leads to the lowest MAE, summed over the five GCM cases. This procedure avoids over-fitting M_{opt} for to any particular GCM case. Since there is seasonal variation in T_{WRF}, we determine a different M_{opt} for each month of the year, resulting in 12 different values of M_{opt}.

Before applying the PODM method, it is a good practice to examine whether the POD basis vectors generated from GCMs in the training dataset can be used to closely approximate a new GCM. If a new GCM cannot be closely approximated by the POD bases from the training dataset, then it's unlikely that PODM will be an effective method for downscaling that GCM. The approximation accuracy can be quantified by determining the *projection error* defined as

$$\varepsilon_{\text{proj}}(\mathbf{p}) = \overline{\mathbf{p}} + \sum_{i=1}^{M}((\zeta_i^{\mathbf{p}})^T \mathbf{p})\zeta_i^{\mathbf{p}} - \mathbf{p}\,, \tag{7}$$

where $\overline{\mathbf{p}}$ and $\zeta^{\mathbf{p}}$ are obtained based on the POD procedure applied to snapshots of $\mathbf{p}$ in the training dataset. For $\mathbf{p} = \Delta\overline{T}_{\mathrm{GCM}}$ of a GCM not in the training dataset, a large $\varepsilon_{\mathrm{proj}}(\mathbf{p})$ implies there are insufficient similarities between the solutions of that GCM and GCMs used in the training data set. It is unlikely in that case that the predictand can be accurately downscaled through the PODM method. In Section 4.1, we use this LOOCV procedure to determine whether $\Delta\overline{T}_{\mathrm{GCM}}$ within the D1 region can be used as a predictor.

3.2 Machine learning based regression approach

In this study, we also consider machine learning (ML) based regression approaches in the statistical downscaling step. State-of-the-art machine learning approaches have been used in ecology (Elith et al. 2008; Maloney et al. 2012; Pittman and Brown 2011) and hydrology (Erdal and Karakurt 2013; Nolan et al. 2015). They are also used to downscale satellite images of land surface temperature (Keramitsoglou et al. 2013). Similar to PODM method, machine-learning algorithms could be advantageous as they are typically automatable and require limited manual tuning. We refer these models as ML-based regression models.

In this paper, different ML methods are used to identify the relationship between the predictors and the predictand. The predictors are defined to be the latitude, longitude, elevation, T_{NARR}, $T_{\mathrm{S,NARR}}$, $\Delta\overline{T}_{\mathrm{GCM}}$, and $\Delta\overline{T}_{\mathrm{S,GCM}}$. ΔT_{WRF} is used as the predictand. This amounts to solving the standard regression problem of relating the predictand to a function of predictors:

$$\Delta T_{\mathrm{WRF}} = f(latitude, longitude, elevation, T_{\mathrm{NARR}}, T_{\mathrm{S,NARR}}, \Delta \overline{T}_{\mathrm{S,GCM}}, \Delta \overline{T}_{\mathrm{S,GCM}})$$

where $f: \mathbb{R}^n \rightarrow \mathbb{R}$, with $n = 7$. Three different machine learning algorithm are used to estimate the function *f*: gradient boosting machines (Freund and Schapire 1997; Friedman et al. 2000; Friedman 2001), extremely randomized trees (Geurts et al. 2006) and elastic net regression method (Zou and Hastie 2005). For further description of these algorithms, see the appendix. Getting the GCM and NARR variables to the WRF resolution is done in two steps. First, the GCM data were interpolated onto NARR grid using bivariate spline approximation on a sphere. Then the GCM and NARR data were interpolated to WRF grid with Gaussian process regression. This combination of preprocessing steps, schematically shown in Figure 2, produced slightly better results than directly interpolating GCM results onto the WRF grid using Gaussian process regression, but the difference is small; our tests showed that the performance of the ML-based regression models only depends weakly on the interpolation scheme.

3.3 Evaluation procedure

To evaluate how well our statistical downscaling methods (SDMs) emulate WRF, we define an approximation error

$$\varepsilon_{\mathrm{SDM}} = T_{\mathrm{SDM}} - T_{\mathrm{WRF,fut}},$$

where SDM can be any of the SDMs used in this study. We will primarily be looking at the mean absolute error (MAE), $e_{\text{MAE,SDM}}$ defined as the average absolute value of ε_{SDM}. Unless otherwise noted, $e_{\text{MAE,SDM}}$ is assumed to be the average over a 10-year simulation while the monthly average $e_{\text{MAE,SDM}}$ is evaluated for a particular month over a 10-year simulation. The error $e_{\text{MAE,SDM}}$ is used to cross-validate the SDMs based on the LOOCV procedure described in Section 3.1.

3.4 Data availability

Temperature output generated by hybrid downscaling with StatWRF (Walton et al. 2016) and with above PODM and ML methods is available from the UCLA Climate Sensitivity Research Lounge website (http://research.atmos.ucla.edu/csrl/pub.html).

4 Results

4.1 Initial analysis of the GCM results

Here do a preliminary check to see if the GCM patterns are similar enough that any pattern can be approximated by POD bases generated from the remaining four GCMs. We determine the MAE of the projection error, $e_{\text{MAE,proj}}$ of $\Delta\overline{T}_{\text{GCM}}$, for each of the five GCMs, when it is left out of the training dataset. We also examine how the projection error depends on the number of POD bases. Figure 3 shows that the $e_{\text{MAE,proj}}$ decreases monotonically with the number of POD basis vectors, M. The GFDL-CM3 case has the highest averaged $e_{\mathrm{MAE,proj}}$, indicating that the GFDL-CM3's $\Delta\overline{T}_{\mathrm{GCM}}$ patterns are least well approximated from the other GCMs. However, for $M = 40$, the mean, and the standard deviation of $e_{\mathrm{MAE,proj}}$ are 0.2 °C and 0.03 °C respectively. This indicates that for a large number of POD bases, PODM can reasonably approximate the left-out GCM pattern, regardless of which GCM is left out. This gives us confidence that PODM is suitable for application for hybrid downscaling, where are the results of a small set of GCMs are extended to a full ensemble.

4.2 Dependence of model accuracy on the predictors and predictand

The accuracy of a statistical model varies based on which combinations of predictors and predictands are used. The two options for predictands are considered: predicting the absolute future temperatures, $T_{\mathrm{WRF,fut}}$, or the difference in temperatures between the future and historical simulations, ΔT_{WRF}. For the predictors there are multiple options: whether to use ΔT_{GCM} or $T_{\mathrm{BC,fut}}$, whether to include surface temperature T_S along with 2m air temperature T, and which domain over which the predictor is sampled. The domain options are the innermost WRF domain (D3) covering the Sierra Nevada, the intermediate WRF domain covering all of California (D2), and the largest WRF domain covering the entire U.S. West Coast and part of the Pacific Ocean (D1; see Figure 1). In each case, only NARR grid cells within the boundaries of the WRF domain are used as the predictor. The GCM data is interpolated onto these NARR grid cells using a bivariate spline interpolation method. PODM is applied to each different combination of predictors, predictand, and domain to determine how each choice affects the resulting statistical model error, $e_{\text{MAE, PODM}}$.

Table 1 shows $e_{\text{MAE, PODM}}$ averaged over all five GCM cases for each combination of choices identified above. In the first three columns, the values in parentheses represent the average $e_{\text{MAE,PODM}}$ over a restricted set of factors. For example, when the predictors are sampled over D1 and ΔT_{WRF} is the predictand, the average $e_{\text{MAE,PODM}}$ over different combinations of remaining factors is 0.45 °C. Table 1 clearly shows that PODM can more accurately predict ΔT_{WRF} than $T_{\text{WRF,fut}}$: the average $e_{\text{MAE,PODM}}$ for ΔT_{WRF} is 0.53 °C, while the average $e_{\text{MAE,PODM}}$ for $T_{\text{WRF,fut}}$ is 0.99 °C. This makes sense as we would expect $T_{\text{WRF,fut}}$ to be a more difficult predictand to approximate using any method, because the time series $T_{\text{WRF,fut}}$ has larger variability due to inclusion of the seasonal cycle, which is not present in ΔT_{WRF}.

Using the largest domain (D1) results in universally greater accuracy compared with the other domains. With ΔT_{WRF} as the predictand, D1 leads to an average $e_{\text{MAE,PODM}}$ that is 12%, and 30% lower than D2, and D3 respectively. This result shows that predictor values outside the predictand domain contain valuable predictive information. However, the predictive value decreases with increasing distance from the predictand domain: the average $e_{\text{MAE,PODM}}$ improves 20% between D3 and D2, and only 12% between D2 and D1.

The inclusion of surface temperature as an auxiliary predictor (i.e. $T_{S,\mathrm{NARR}}$ and $\Delta\overline{T}_{S,\mathrm{GCM}}$) also universally improves the accuracy of the models. When $T_{\mathrm{WRF,fut}}$ is the predictand, this inclusion reduces the error by 11-17%, depending on the domain size of the predictor. However, when ΔT_{WRF} is the predictand, the improvement is smaller: the error is only reduced by 6.5% when the predictor domain is D1 and no reduction is observed for the predictor domain of D2 and D3. Finally, the alternative formulation of the predictors to mimic boundary conditions ($T_{\mathrm{BC,fut}}$) improves accuracy when $T_{\mathrm{WRF,fut}}$ is the predictand, and decreases accuracy when ΔT_{WRF} is the predictand ($e_{\mathrm{MAE,PODM}}$ is larger by up to 6%). This makes sense, as one would generally expect best results when the form of the predictor matches the form of the predictand.

The above analysis guides our formulation of PODM and other SDMs in the following sections. First, results clearly indicate that the statistical model should be trained to predict ΔT_{WRF}, as opposed to predicting $T_{\mathrm{WRF,fut}}$, regardless of whether the goal is to predict ΔT_{WRF} or $T_{\mathrm{WRF,fut}}$. So, we formulate PODM and the SDMs with ΔT_{WRF} as the predictand. We use domain D1, as it leads to a higher accuracy model when compared to the alternatives. $T_{S,\mathrm{NARR}}$ and $\Delta\overline{T}_{S,\mathrm{GCM}}$ are included as predictors since they slightly improve PODM accuracy and the additional computational cost is minimal. Finally, we use $\Delta\overline{T}_{\mathrm{GCM}}$ instead of $T_{\mathrm{BC,fut}}$ as predictor, since it better matches the form of our predictand ΔT_{WRF}.

4.3 Comparison of statistical methods in approximating WRF

The $e_{\mathrm{MAE,SDM}}$ of the different SDMs for all the GCM cases are shown in Table 2. The average error for PODM, $e_{\mathrm{MAE,PODM}}$, over the five GCMs is 0.44 °C. The monthly average $e_{\mathrm{MAE,PODM}}$ varies with month, and the variation is different for each of the GCM cases (Figure 4). However, when averaged over the five GCM cases, the monthly average $e_{\mathrm{MAE,PODM}}$ varies more smoothly, with higher errors in the summer months (reaching a maximum in July) and lower errors in the winter months (reaching a minimum in December). ML-based regression models are universally less accurate than the PODM model (Table 2). The average $e_{\mathrm{MAE,SDM}}$ of the different ML-based regression models are 50%–110% larger than for PODM model.

PODM also more accurately captures changes in climatology, $\Delta\bar{T}_{\mathrm{WRF}}$ (Table 3). The MAEs for the 10-year monthly temperature climatology ($e^{\Delta\bar{T}}_{\mathrm{MAE,SDM}}$) of the ML-based regression models are larger than that of PODM model by 45%–130% (Table 3). StatWRF errors are 21% larger than PODM. PODM, like StatWRF, has much lower errors compared to traditional statistical downscaling methods as well, including Bias Correction and Constructed Analogs (BCCA; Maurer and Hidalgo 2008) and Bias Correction with Spatial Disaggregation (BCSD; Wood et al. 2004). (Downscaled CMIP5 climate projections using BCCA and BCSD were obtained from http://gdo-dcp.ucllnl.org/downscaled_cmip_projections/, Reclamation 2013.) Interestingly, ML-based regression models do not perform any better than these well-established downscaling techniques despite having higher degrees of complexity. Given the poor performance of ML-based regression models, we focus only on PODM models in subsequent sections.

4.4 Comparing spatial distributions of the predictions

Here, we compare the spatial patterns of $\Delta\overline{T}_{\mathrm{PODM}}$, $\Delta\overline{T}_{\mathrm{WRF}}$, and $\Delta\overline{T}_{\mathrm{GCM}}$. Figure 5 shows January and July as examples of months where $\Delta\overline{T}_{\mathrm{PODM}}$ poorly and closely matches $\Delta\overline{T}_{\mathrm{WRF}}$, respectively. In January, there is a large disparity between the GCM temperature changes and the WRF-downscaled temperature changes. Importantly, these biases are not in the same direction for CNRM-CM5 and GFDL-CM3. WRF-downscaled CNRM-CM5 has much less warming than CNRM-CM5 (about 1–2 °C). Meanwhile, WRF-downscaled GFDL-CM3 has much *more* warming than GFDL-CM3 (about 1–2 °C). When relationship between the WRF-downscaled warming and GCM warming is inconsistent between the cases, it is challenging for any statistical model to accurately model it. Thus, PODM struggles to predict WRF-downscaled temperatures in January. PODM better predicts WRF in July, when the relationship between the GCM warming and the WRF-downscaled warming is more consistent.

We now compare changes in temperature climatology averaged over the five GCM cases, denoted as $\langle \Delta\overline{T}_{\mathrm{PODM}} \rangle$, $\langle \Delta\overline{T}_{\mathrm{GCM}} \rangle$, and $\langle \Delta\overline{T}_{\mathrm{WRF}} \rangle$. Figure 6 shows that differences between $\langle \Delta\overline{T}_{\mathrm{PODM}} \rangle$ and $\langle \Delta\overline{T}_{\mathrm{WRF}} \rangle$ are typically small (< 0.5 °C) except for the month of June. PODM is able to capture the fine-scale details present in $\langle \Delta\overline{T}_{\mathrm{WRF}} \rangle$, such as snow albedo feedback (Walton et al. 2016). This demonstrates that an automated, objective statistical model can capture important features that had to be parameterized in previous hybrid downscaling attempts.

4.5 Downscaling 35 CMIP5 GCMs

The purpose of hybrid downscaling is to enable rapid, high-quality downscaling of output from a large number of GCMs. Here we demonstrate this capability by applying PODM to 35 CMIP5 GCMs run under the RCP8.5 forcing scenario. Before applying PODM, we check whether the original five GCMs are good representatives of the full ensemble of GCMs. If so, then we can have more confidence that the PODM will have similar accuracy in downscaling the new GCMs as it does in downscaling the original five. To do this, we approximate the full ensemble of GCM warming patterns using POD bases constructed from the five original GCM warming patterns, similar to the analysis in Section 4.1. For the full ensemble, the mean and standard deviation of the approximation errors $e_{\mathrm{MAE,proj}}$ for are 0.24 °C and 0.07 °C, respectively. In comparison, the mean and standard deviation are 0.2 °C and 0.03 °C for the LOOCV errors obtained with the original five GCMs. Since these values are of similar magnitudes, we expect PODM to emulate WRF to a similar degree of accuracy as was found in Section 4.3 when downscaling the entire ensemble.

To downscale the full ensemble, PODM is trained on data from all six WRF simulations (1 historical + 5 future) as described in Section 2. For some GCMs, surface temperature output is not available in the CMIP5 database. For these GCMs, surface temperature changes are not included as a predictor. This should not significantly alter the accuracy as including surface temperature resulted in only minimal gains (6.5% improvement). Figure 7 shows that PODM captures spatial variations due to snow albedo feedback and the complex topography of the Sierra Nevada that are visible in the WRF solution, but not in the original GCM data.

5 Discussion

Our results show that PODM can emulate WRF in downscaling temperature changes with errors less than 0.44 °C. This is slightly higher accuracy than the original StatWRF model proposed by Walton et al. (2016). Additionally, PODM is objective and the training steps can be automated. In contrast, StatWRF requires that the user parameterizes salient physical processes affecting the climate change signal in the region of interest and to manually determine the appropriate large-scale predictors. Thus PODM can be applied quickly to any region, without the user needing expert knowledge about the region's climate. We note that the skill of any statistical model — including PODM — in emulating WRF is likely to be region dependent, so hybrid downscaling users need to verify PODM's skill when applying it elsewhere. It is also important to acknowledge that while the PODM model can be valuable tool for making downscaled projections, it may not enhance our understanding of the climate processes at play. For instance, the way PODM model accounts for the additional warming in the Sierra Nevada due to snow albedo feedback is part of PODM's internal workings that is opaque to the user.

The skill of PODM could be improved if more dynamically downscaled GCM simulations are included in the training data. Indeed, the PODM model described in this paper could achieve even higher accuracy if the dynamically downscaled GCMs are chosen to best represent the full set of GCMs. The five GCM cases currently used in this paper were chosen to represent the range of temperature and precipitation changes predicted by the GCMs (Walton et al. 2016). However, if just temperature projections are desired, a more representative set of GCMs could be selected by identifying the five GCMs that minimize $e_{\mathrm{MAE,proj}}$ when their POD bases are used to approximate the rest of the ensemble of GCMs. We will study these training procedures in our future work.

It's important to acknowledge that accuracy results and optimal predictor/predictand combinations for PODM might change in downscaling studies that do not use the pseudo-global warming methodology (PGW). In our study, each future WRF simulation is downscaling of historical NARR plus a change in GCM climatology. Thus, interannual variability of each future simulation is nearly identical to the historical simulation. So, when subtracting the future and historical sequences to determine ΔT_{WRF}, the interannual variability mostly cancels. In contrast, if raw GCM output were downscaled for the future and historical simulations, as is often the case, then interannual variability between the historical and future simulations will be unrelated. In this case, ΔT_{WRF} could have considerably more variability that PODM needs to capture, and accuracy could be lower. Furthermore, different formulation of the predictors will be needed. The historical and future GCM sequences $T_{\text{GCM,hist}}$ and $T_{\text{GCM,fut}}$ will probably need to be included, not just the GCM difference in climatology, $\Delta \bar{T}_{\text{GCM}}$.

6 Conclusions

In this study, we have demonstrated that an objective statistical downscaling method, PODM, can approximate WRF temperatures changes with similar or better accuracy than previous statistical methods in California's Sierra Nevada mountain range. ML-based regression methods were also tested, but were found to be much less accurate than PODM in emulating WRF. Our analysis shows that use of a large predictor domain encompassing the entire U.S. West Coast yielded the highest accuracy, even though our predictand domain is limited to the Sierra Nevada mountain range. The inclusion of surface temperature as an additional predictor was found to moderately improve accuracy. Our results also show that if the goal is to predict future temperatures, then statistical models should be designed to predict them as anomalies from the current climate (as opposed to directly predicting the future values).

Acknowledgement

George Pau and Samir Touzani were supported by the Director, Office of Science, Office of Biological and Environmental Research of the US Department of Energy under Contract #DEAC02-05CH11231 as part of the Early Career Research Program. Daniel Walton was supported by the US Department of Energy (Grant #DE-SC0014061, "Developing Metrics to Evaluate the Skill and Credibility of Downscaling"). The authors are not aware of any conflicts of interest.

This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the US Department of Energy under Contract #DEAC02-05CH11231.

Appendix

In this appendix, gradient boosting machine, Extra-Tree and Elastic Net are reviewed. The first two are so-called ensemble machine learning algorithms based on regression trees and the latter is a linear regression method using a combination of L1 and L2 penalties. Hybrid downscaled output using these methods are available from the UCLA Climate Sensitivity Research Lounge website http://research.atmos.ucla.edu/csrl/pub.html.

A.1 Gradient boosting machine (GBM)

The gradient boosting machine (GBM) algorithm (Freund and Schapire 1997; Friedman et al. 2000; Friedman 2001) combines iteratively several simple models, called “weak learners”, in order to obtain a “strong learner” with improved prediction accuracy. GBM starts by initializing the model by a first guess of a regression tree model (Breiman et al. 1984) that maximally reduces the loss function (i.e. least squares). Then at each step a new regression tree model is fitted to the current residual and added to the previous model in order to update the residual, until the number of iteration K is reached. By fitting the regression tree model to the residuals the global model is improved in the regions where it is not accurate.

GBM expresses the relationship between the scalar predictand (corresponding to ΔT_{WRF}) and the n scalar predictors ($\boldsymbol{p} = \{p^1, \ldots, p^n\}$, corresponding to the latitude, longitude, elevation, T_{NARR}, $T_{\mathrm{S,NARR}}$, $\Delta \overline{T}_{\mathrm{GCM}}$, and $\Delta \overline{T}_{\mathrm{S,GCM}}$) as an ensemble of K additive functions:

$$f_{\mathrm{GBM}}(\boldsymbol{p}) = \sum_{k=1}^{K} \phi_k(\boldsymbol{p}), \tag{8}$$

where $\phi_k(\boldsymbol{p})$ is a regression tree model. Note that K is the number of GBM iteration steps and $\boldsymbol{p}$ represents m different scalar predictors at a particular grid block while **p** represents a multivariate predictor.

As for any predictive machine learning algorithm, GBM has several parameters that need to be tuned. These parameters are: 1) the complexity of the regression tree, which is represented by the maximum number of split points of the decision tree; 2) K the number of the algorithm iterations; 3) the learning rate, which is a relatively small positive value between 0 and 1, and inversely proportional to K; and 4) the fraction of training data that is used as a training subsample at each iterative step. To choose the combination of these parameters that produce the best predictive GBM model the LOOCV procedure (described in Section 3.1) combined with the so-called search grid method have been used in this study. This method is based on predefining a grid of GBM parameters combinations, then for each combination a GBM model is estimated. The best combination is selected as the one that produce the most accurate model using the LOOCV procedure. We used the minimization of the MAE as the accuracy criteria to select the best combination.

The XGBoost python library (https://github.com/dmlc/xgboost), which is a relatively new efficient implementation of GBM method, is used here. The performance of XGBoost has been demonstrated in multiple data mining and machine learning challenges (Chen and Guestrin 2016). We refer readers to Chen and Guestrin (2016) for details of the XGBoost algorithm, especially the advanced features that have been implemented in it.

A.2 Extremely Randomized Trees (Extra-Trees)

Similar to GBM the Extra-Tree algorithm (Geurts et al. 2006) is based on a simple averaging of the weak learner while the boosting algorithm of GBM is built upon a constructive iterative strategy. It builds a set of regression trees, which are trained by selecting the decision trees splits points at random. In other words, instead of selecting the splits points that are locally optimal, these splits points are selected randomly. The predictions of each regression trees are simply averaged to create the final prediction.

The Extra-Tree procedure has two main parameters that need to be tuned. These parameters are the maximum number of splits points of each regression tree and K the number of regression trees of the ensemble. As for the GBM model the best combination is selected using a search grid and the previously described LOOCV procedure. In this work, we have used the Extra-Trees implementation of the scikit-learn python library (Pedregosa et al. 2011).

A.3 Elastic net linear regression

We consider the standard linear regression model, which is defined for the given *n* scalar predictors $\boldsymbol{p} = \{p^1, \ldots, p^n\}$ (corresponding to the latitude, longitude, elevation, T_{NARR}, $T_{\mathrm{S,NARR}}$, $\Delta\overline{T}_{\mathrm{GCM}}$, and $\Delta\overline{T}_{\mathrm{S,GCM}}$) and the scalar predictand f (corresponding to ΔT_{WRF}) by:

$$f_{\mathrm{linear}}(\boldsymbol{p}) = \sum_{i=1}^{n} \beta_i p^i,$$

The standard approach to estimate the regression coefficients $\beta = \{\beta_1, \ldots, \beta_n\}$ is to use the ordinary least squares algorithm (OLS). However it is well known that the OLS often underperforms in term of prediction accuracy compared to other linear techniques such as ridge regression (Hoerl and Kennard 2000) and LASSO (Tibshirani 1996). The first one applies an L2 penalty on the coefficients and the second one applies an L1 penalty.

Elastic Net regression method (Elastic Net, Zou and Hastie 2005) applies a convex combination of L1 and L2 penalties on the regression coefficients. There are two parameters to optimize: the ratio of L1 penalty to L2 penalty, and the magnitude of the total penalty. These two parameters are tuned using the same methodology as for GBM and Extra-Tree algorithms. The scikit-learn python library (Pedregosa et al. 2011) implementation of the Elastic Net has been used in this work.

Reference

Benestad, R. E., I. Hanssen-Bauer, and D. Chen, 2008: *Empirical-statistical downscaling.* World Scientific.

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. *Journal of Climate*, **28,** 401-421.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen, 1984: *Classification and regression trees.* CRC press.

Chen, T., and C. Guestrin, 2016: XGBoost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*.

Elith, J., J. R. Leathwick, and T. Hastie, 2008: A working guide to boosted regression trees. *Journal of Animal Ecology*, **77,** 802-813.

Erdal, H. I., and O. Karakurt, 2013: Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms. *J Hydrol*, **477,** 119-128.

Everson, R., and L. Sirovich, 1995: Karhunen–Loeve procedure for gappy data. *Journal of the Optical Society of America A*, **12,** 1657-1664.

Fowler, H. J., S. Blenkinsop, and C. Tebaldi, 2007: Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology*, **27,** 1547-1578.

Freund, Y., and R. E. Schapire, 1997: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, **55,** 119-139.

Friedman, J., T. Hastie, and R. Tibshirani, 2000: Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, **28,** 337-407.

Friedman, J. H., 2001: Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, **29,** 1189-1232.

Geurts, P., D. Ernst, and L. Wehenkel, 2006: Extremely randomized trees. *Mach Learn*, **63,** 3-42.

Gosling, S. N., and Coauthors, 2011: A review of recent developments in climate change science. Part II: The global-scale impacts of climate change. *Progress in Physical Geography*, **35,** 443-464.

Gutmann, E., T. Pruitt, M. P. Clark, L. Brekke, J. R. Arnold, D. A. Raff, and R. M. Rasmussen, 2014: An intercomparison of statistical downscaling methods used for water resource assessments in the United States. *Water Resour Res*, **50,** 7167-7186.

Hanssen-Bauer, I., E. J. Forland, J. E. Haugen, and O. E. Tveito, 2003: Temperature and precipitation scenarios for Norway: comparison of results from dynamical and empirical downscaling. *Climate Research*, **25,** 15-27.

Hoerl, A. E., and R. W. Kennard, 2000: Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **42,** 80-86.

Keramitsoglou, I., C. T. Kiranoudis, and Q. Weng, 2013: Downscaling Geostationary Land Surface Temperature Imagery for Urban Analysis. *IEEE Geoscience and Remote Sensing Letters*, **10,** 1253-1257.

Maloney, K. O., M. Schmid, and D. E. Weller, 2012: Applying additive modelling and gradient boosting to assess the effects of watershed and reach characteristics on riverine assemblages. *Methods in Ecology and Evolution*, **3,** 116-128.

Maraun, D., and Coauthors, 2010: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, **48**.

Maurer, E. P., and H. G. Hidalgo, 2008: Utility of daily vs. monthly large-scale climate data: an intercomparison of two statistical downscaling methods. *Hydrol. Earth Syst. Sci.*, **12,** 551-563.

McCarthy, M. P., M. J. Best, and R. A. Betts, 2010: Climate change in cities due to global warming and urban effects. *Geophys Res Lett*, **37,** n/a-n/a.

Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bulletin of the American Meteorological Society*, **87,** 343-360.

Niu, G. Y., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J Geophys Res-Atmos*, **116**.

Nolan, B. T., M. N. Fienen, and D. L. Lorenz, 2015: A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA. *J Hydrol*, **531, Part 3,** 902-911.

Pau, G. S. H., G. Bisht, and W. J. Riley, 2014: A reduced-order modeling approach to represent subgrid-scale hydrological dynamics for land-surface simulations: application in a polygonal tundra landscape. *Geosci. Model Dev.*, **7,** 2091-2105.

Pau, G. S. H., C. Shen, W. J. Riley, and Y. Liu, 2016: Accurate and efficient prediction of fine-resolution hydrologic and carbon dynamic simulations from coarse-resolution models. *Water Resour Res*, **52,** 791-812.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12,** 2825-2830.

Pittman, S. J., and K. A. Brown, 2011: Multi-scale approach for predicting fish species distributions across coral reef seascapes. *Plos One*, **6,** e20583.

Reclamation, 2013: 'Downscaled CMIP3 and CMIP5 Climate and Hydrology Projections: Release of Downscaled CMIP5 Climate Projections, Comparison with preceding Information, and Summary of User Needs', prepared by the U.S. Department of the Interior, Bureau of Reclamation, Technical Services Center, Denver, Colorado. 47pp.

Riahi, K., and Coauthors, 2011: RCP 8.5—A scenario of comparatively high greenhouse gas emissions. *Climatic Change*, **109,** 33.

Robinson, T., M. Eldred, K. Willcox, and R. Haimes, 2006: Strategies for Multifidelity Optimization with Variable Dimensional Hierarchical Models. *47th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, American Institute of Aeronautics and Astronautics.

Skamarock, W. C., and Coauthors, 2008: A description of the advanced research WRF Ver. 30. NCAR Technical Note.

Sun, A. Y., R. M. Miranda, and X. L. Xu, 2015a: Development of multi-metamodels to support surface water quality management and decision making. *Environ Earth Sci*, **73,** 423-434.

Sun, F., D. B. Walton, and A. Hall, 2015b: A Hybrid Dynamical–Statistical Downscaling Technique. Part II: End-of-Century Warming Projections Predict a New Climate State in the Los Angeles Region. *Journal of Climate*, **28,** 4618-4636.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, **93,** 485-498.

Tebaldi, C., and J. Arblaster, 2014: Pattern scaling: Its strengths and limitations, and an update on the latest model simulations. *Climatic Change*, **122,** 459-471.

Tibshirani, R., 1996: Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met*, **58,** 267-288.

von Storch, H., E. Zorita, and U. Cubasch, 1993: Downscaling of Global Climate Change Estimates to Regional Scales: An Application to Iberian Rainfall in Wintertime. *Journal of Climate*, **6,** 1161-1171.

Walton, D., A. Hall, F. Sun, M. Schwartz, and N. Berg, 2016: Incorporating Snow Albedo Feedback into Downscaled Temperature and Snow Cover Projections for California's Sierra Nevada. *Journal of Climate*, **30**, 1417-1438, doi: http://dx.doi.org/10.1175/JCLI-D-16-0168.1.

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. *Journal of Climate*, **28,** 4597-4617.

Wilby, R. L., T. M. L. Wigley, D. Conway, P. D. Jones, B. C. Hewitson, J. Main, and D. S. Wilks, 1998: Statistical downscaling of general circulation model output: A comparison of methods. *Water Resour Res*, **34,** 2995-3008.

Wood, A. W., L. R. Leung, V. Sridhar, and D. P. Lettenmaier, 2004: Hydrologic Implications of Dynamical and Statistical Approaches to Downscaling Climate Model Outputs. *Climatic Change*, **62,** 189-216.

Zou, H., and T. Hastie, 2005: Regularization and variable selection via the elastic net. *J Roy Stat Soc B*, **67,** 301-320.

Table 1: Averaged $e_{\mathrm{MAE,PODM}}$ for different combinations predictors and predictands. The value in parenthesis is the averaged $e_{\mathrm{MAE,PODM}}$ over that particular parameter.

Predictand (°C)	Predictor: Domain (°C)	Predictor: $T_{S,\mathrm{NARR}}$ and $\Delta\overline{T}_{S,\mathrm{GCM}}$ (°C)	Predictor: $\Delta\overline{T}_{\mathrm{GCM}}$ or $T_{\mathrm{BC,fut}}$	$e_{\mathrm{MAE,PODM}}$, °C
T_{WRF} (0.99)	D01 (0.68)	False (0.73)	$\Delta\overline{T}_{\mathrm{GCM}}$	0.77
			$T_{\mathrm{BC,fut}}$	0.69
		True (0.62)	$\Delta\overline{T}_{\mathrm{GCM}}$	0.63
			$T_{\mathrm{BC,fut}}$	0.61
	D02 (0.96)	False (1.05)	$\Delta\overline{T}_{\mathrm{GCM}}$	1.04
			$T_{\mathrm{BC,fut}}$	1.05
		True (0.87)	$\Delta\overline{T}_{\mathrm{GCM}}$	0.90
			$T_{\mathrm{BC,fut}}$	0.84
	D03 (1.34)	False (1.42)	$\Delta\overline{T}_{\mathrm{GCM}}$	1.44
			$T_{\mathrm{BC,fut}}$	1.39
		True (1.26)	$\Delta\overline{T}_{\mathrm{GCM}}$	1.37
			$T_{\mathrm{BC,fut}}$	1.15
ΔT_{WRF}	D01	False	$\Delta\overline{T}_{\mathrm{GCM}}$	0.45

(0.53)	(0.45)	(0.46)	$T_{\mathrm{BC,fut}}$	0.46
		True (0.43)	$\Delta\overline{T}_{\mathrm{GCM}}$	0.43
			$T_{\mathrm{BC,fut}}$	0.43
	D02 (0.51)	False (0.51)	$\Delta\overline{T}_{\mathrm{GCM}}$	0.49
			$T_{\mathrm{BC,fut}}$	0.52
		True (0.51)	$\Delta\overline{T}_{\mathrm{GCM}}$	0.50
			$T_{\mathrm{BC,fut}}$	0.52
	D03 (0.64)	False (0.64)	$\Delta\overline{T}_{\mathrm{GCM}}$	0.64
			$T_{\mathrm{BC,fut}}$	0.65
		True (0.64)	$\Delta\overline{T}_{\mathrm{GCM}}$	0.62
			$T_{\mathrm{BC,fut}}$	0.66

Table 2: The MAEs ($e_{\mathrm{MAE,SDM}}$) of ΔT_{SDM} for the different SDMs, for each GCM cases, and the ensemble averages.

GCM	**PODM**	**Elastic Net**	**GBM**	**Extra-Tree**
CNRM-CM5	0.40	0.86	0.80	0.87
GFDL-CM3	0.54	0.83	0.94	1.26

INMCM4	0.38	0.46	0.97	0.90
IPSL-CM5A-LR	0.48	0.72	1.14	1.15
MPI-ESM-LR	0.38	0.44	0.50	0.51
Average	**0.44**	**0.66**	**0.87**	**0.94**

Table 3: The MAEs ($e^{\Delta\overline{T}}_{\mathrm{MAE,SDM}}$) of $\Delta\overline{T}_{\mathrm{SDM}}$ for the different SDMs, for each GCM cases, and the ensemble averages. BCCA stands for Bias Correction and Constructed Analogs and BCSD stands for Bias Correction with Spatial Disaggregation. MAE data for StatWRF, BCCA, BCSD, and linter interpolation are from Walton et al. (2016).

GCM	**PODM**	**Elastic Net**	**GBM**	**Extra-Tree**	**Stat-WRF**	**BCCA**	**BCSD**	**Linear inter-polation**
CNRM-CM5	0.34	0.76	0.66	0.76	0.52	0.49	0.89	0.85
GFDL-CM3	0.47	0.66	0.80	1.25	0.61	1.18	1.08	0.75
INMCM4	0.31	0.34	0. 88	0.84	0.47	0.91	0.94	0.48
IPSL-CM5A-LR	0.41	0.68	1.10	1.11	0.31	0.78	0.56	0.43
MPI-ESM-LR	0.29	0.24	0. 28	0.30	0.35	0.63	0.43	0.44

Average	0.37	0.54	0. 74	0.85	0.45	0.80	0.78	0.59

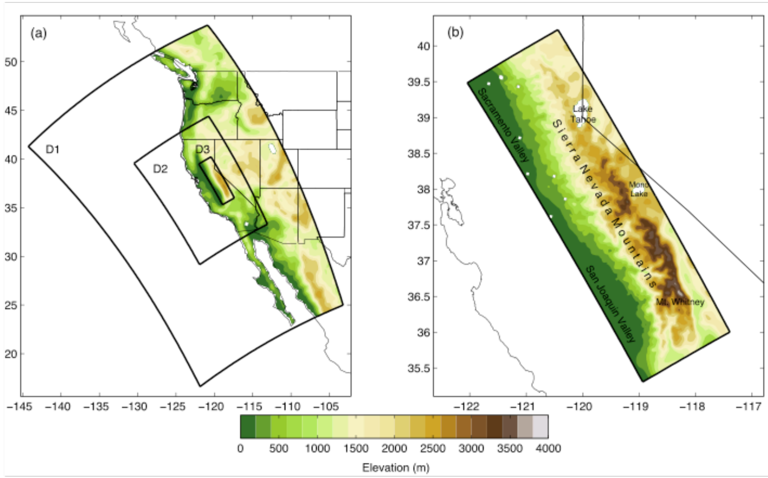

Figure 1: (Taken from Walton et al., 2016) (a) Elevation (meters) and model setup with three one-way nested WRF domains (D1, D2, and D3) at horizontal resolutions of 27, 9, and 3 km. (b) Innermost domain elevation (meters).

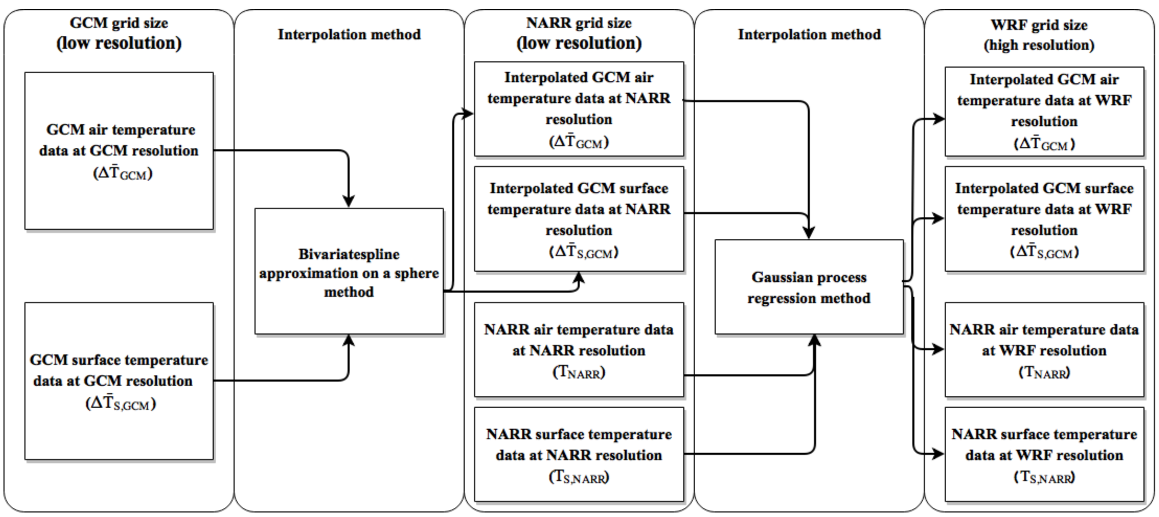

Figure 2: Predictors data pre-processing steps for the downscaling approach based on ML-based regression models.

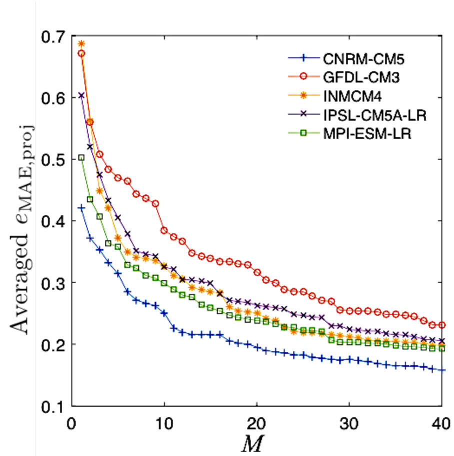

Figure 3: The projection error $e_{\mathrm{MAE,proj}}$ for approximating the left-out $\Delta\bar{T}_{\mathrm{GCM}}$ pattern, when M bases are used. Here, $e_{\mathrm{MAE,proj}}$ is an average over all 12 calendar months.

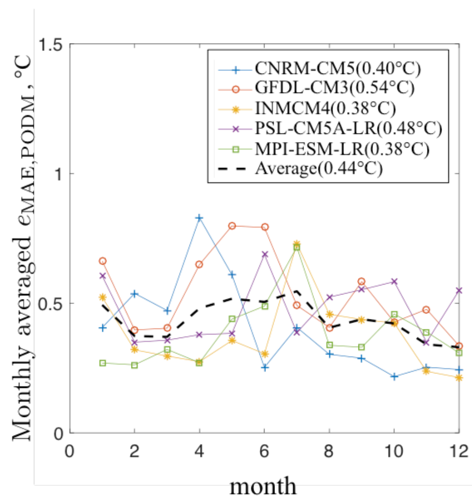

Figure 4: Monthly averaged $e_{\mathrm{MAE,PODM}}$ versus month for different GCM cases. The values in the parentheses are the averages for each of the GCM cases.

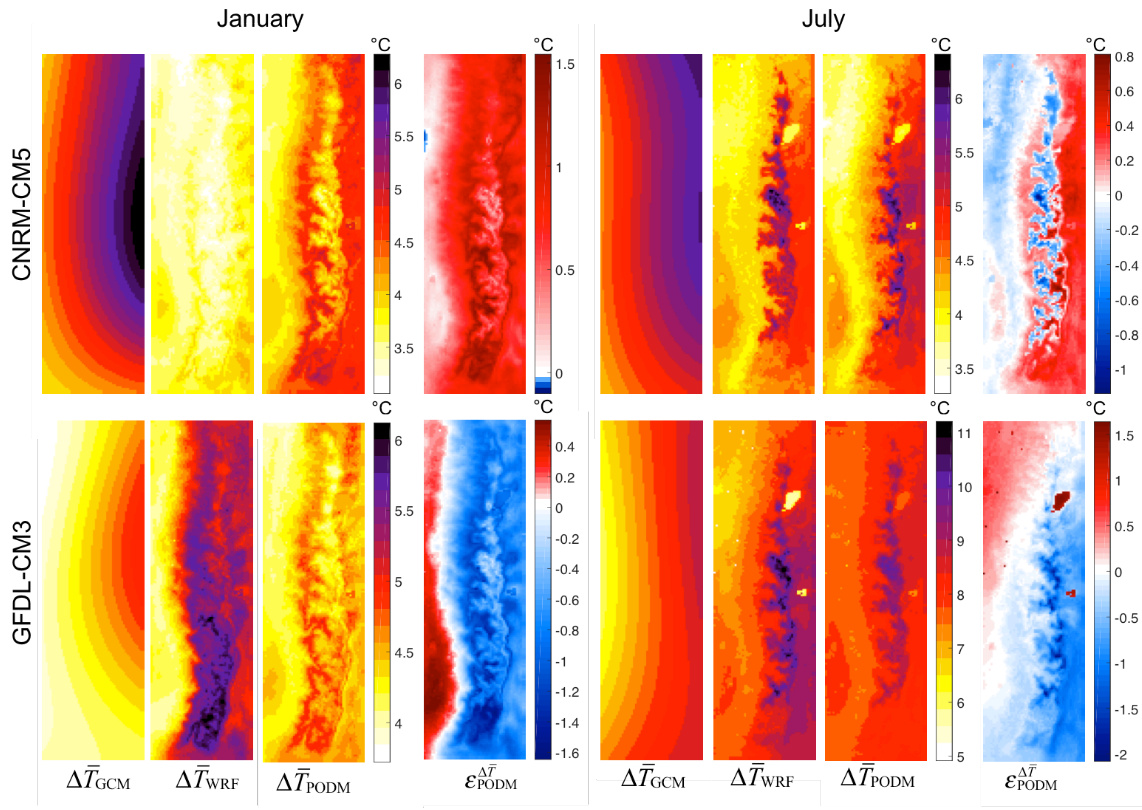

Figure 5: $\Delta\overline{T}_{\mathrm{GCM}}$, $\Delta\overline{T}_{\mathrm{WRF}}$, $\Delta\overline{T}_{\mathrm{PODM}}$, and $\varepsilon^{\Delta\overline{T}}_{\mathrm{PODM}} = \Delta\overline{T}_{\mathrm{PODM}} - \Delta\overline{T}_{\mathrm{WRF}}$ in January and July. Only results from the CNRM-CM5 and GFDL-CM3 cases are shown.

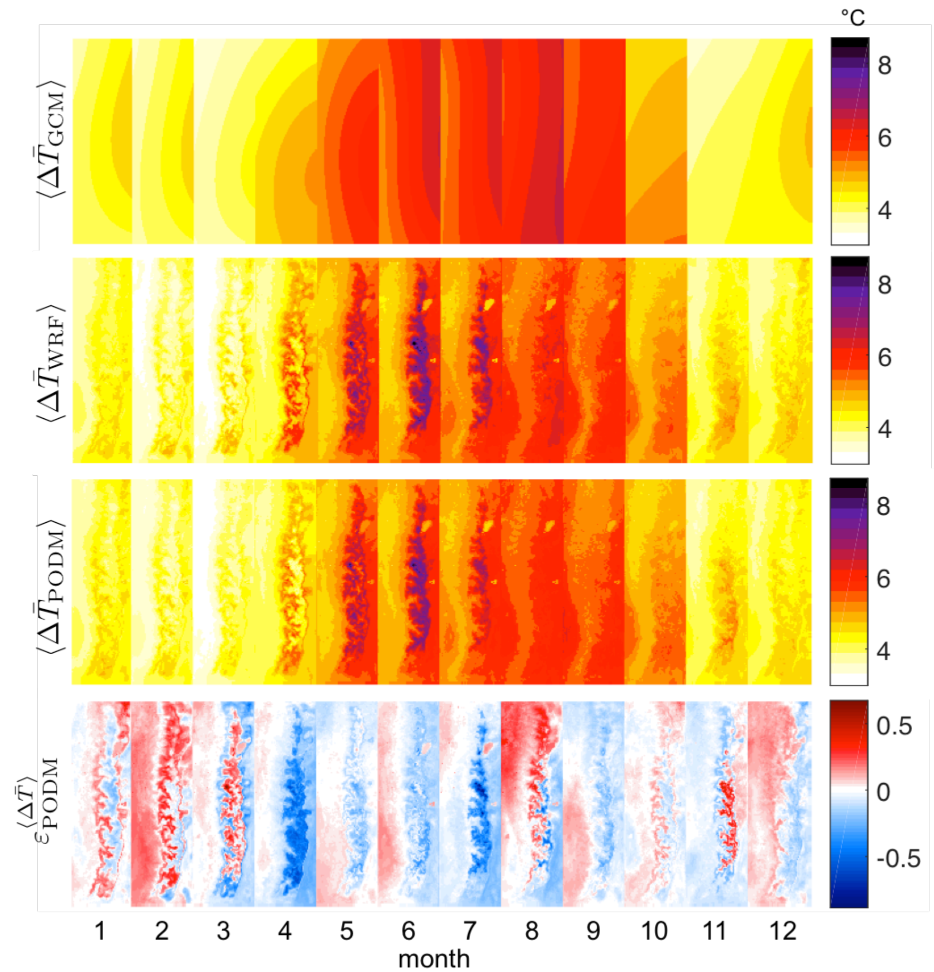

Figure 6: Changes in temperature climatology averaged over five GCM cases produced from three sources. (row 1) GCM changes $\langle \Delta \bar{T}_{\mathrm{GCM}} \rangle$. (row 2) WRF changes $\langle \Delta \bar{T}_{\mathrm{WRF}} \rangle$. (row 3) PODM changes produced via cross validation $\langle \Delta \bar{T}_{\mathrm{PODM}} \rangle$. (row 4) differences between WRF and PODM $\varepsilon_{\mathrm{PODM}}^{\langle \Delta \bar{T} \rangle}$ for each of month, averaged over the five GCMs cases.

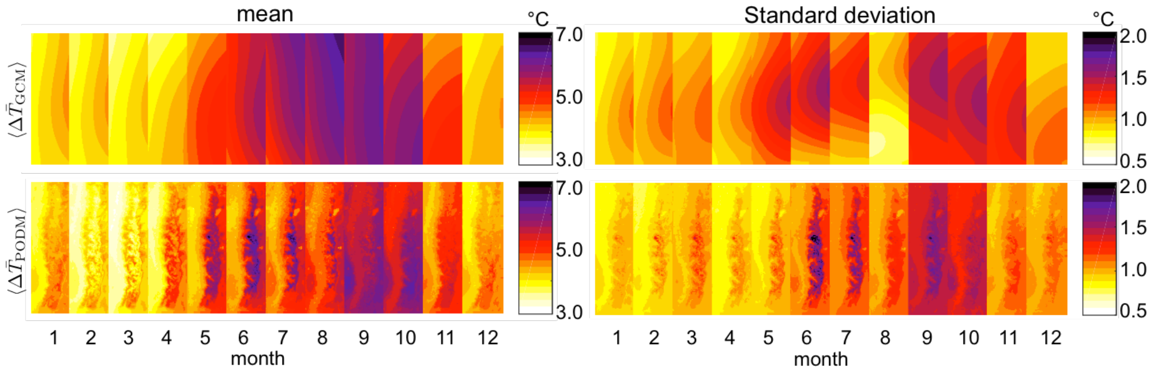

Figure 7: The mean and the standard deviation of the monthly $\Delta\overline{T}_{\mathrm{PODM}}$ compared to $\Delta\overline{T}_{\mathrm{GCM}}$.