JGR Oceans

RESEARCH ARTICLE
10.1029/2020JC017140

Key Points:

- Bottom temperature and salinity simulations from ocean circulation models often have spatially and temporally autocorrelated biases
- Regression kriging can reduce these biases while preserving spatial-temporal model trends
- The Structural Similarity Index can assess the accuracy, precision, and spatial similarities of model output and bias correction methods

Correspondence to:
J.-H. Chang,
jui-han.chang@noaa.gov

Citation:
Chang, J.-H., Hart, D. R., Munroe, D. M., & Curchitser, E. N. (2021). Bias correction of ocean bottom temperature and salinity simulations from a regional circulation model using regression kriging. *Journal of Geophysical Research: Oceans*, *126*, e2020JC017140. https://doi.org/10.1029/2020JC017140

Received 8 JAN 2021
Accepted 20 MAR 2021

Bias Correction of Ocean Bottom Temperature and Salinity Simulations From a Regional Circulation Model Using Regression Kriging

Jui-Han Chang[1], Deborah R. Hart[1], Daphne M. Munroe[2], and Enrique N. Curchitser[3]

[1]Northeast Fisheries Science Center, NMFS/NOAA, Woods Hole, MA, USA, [2]Haskin Shellfish Research Laboratory, Rutgers University, Port Norris, NJ, USA, [3]Department of Environmental Science, Rutgers University, New Brunswick, NJ, USA

Abstract It is well known that climate and circulation model simulation output are often systematically biased. However, existing bias correction methods typically ignore spatial autocorrelation of the biases and correct only the overall mean and variance, resulting in mismatched spatial variability between bias-corrected simulations and observations. In this study, we propose using regression kriging (RK) to correct for biased spatial patterns and apply this method to Regional Ocean Modeling System (ROMS) simulated ocean bottom temperature and salinity for the Mid-Atlantic Bight, USA. RK combines modeling non-stationary trends using (generalized) regression with ordinary kriging (OK) of the regression residuals. We compared the performance of RK to a simpler OK method and a quantile mapping (QM) method often used for bias correction of such model output. These methods were evaluated using the Structural Similarity (SSIM) index that can simultaneously evaluate model accuracy, precision, and spatial similarities. Our results show that while both RK and QM can correct for overall biases of the mean and variation, only RK can effectively reduce the spatial-temporal autocorrelation of the biases. The RK method was able to bias correct while preserving the spatial-temporal trends of the ROMS simulated bottom temperature and salinity surfaces. The RK approach can easily be applied to any similar climate and circulation model simulation output. This study has profound implications for studies that use the output from such a model for fine-scale mapping, habitat suitability modeling, species distribution modeling, or predicting the effects of climate change.

Plain Language Summary Climate and circulation model predictions often differ systematically from direct observations (i.e., are biased), and these differences commonly show spatial patterns. Existing bias correction methods typically correct only the spatially averaged biases but not for spatial patterns in these biases. In this study, we propose using regression kriging (RK) to correct for biased spatial patterns and apply this method to Regional Ocean Modeling System (ROMS) simulated ocean bottom temperature and salinity for the Mid-Atlantic Bight, USA. We compared the performance of RK to a quantile mapping (QM) method often used for bias correction of such model output. These methods were evaluated using the Structural Similarity (SSIM) index that can simultaneously evaluate model accuracy, precision, and spatial similarities. Our results show that while both RK and QM can correct for spatially averaged biases, only RK can effectively improve the spatial similarity of the ROMS simulated and observed bottom temperatures and salinity surfaces. The RK approach can easily be applied to any similar climate and circulation model simulation output. This work has profound implications for studies that use the output from such a model for detailed mapping or determining habitat suitability, species distribution, and the effects of climate change.

1. Introduction

Climate and circulation model simulations estimate climatological and environmental variables continuously across space from the past to present and sometimes into the future. The continuous nature of these estimates provides valuable information that is essential for assessing the impacts of climatological and environmental changes. However, it is well known that these simulation outputs are often systematically biased relative to observations at spatial and temporal scales of interest, and require bias corrections before further use (Aung et al., 2016; Giorgi, 2019; Maraun, 2016; Shrestha et al., 2017; Teutschbein &

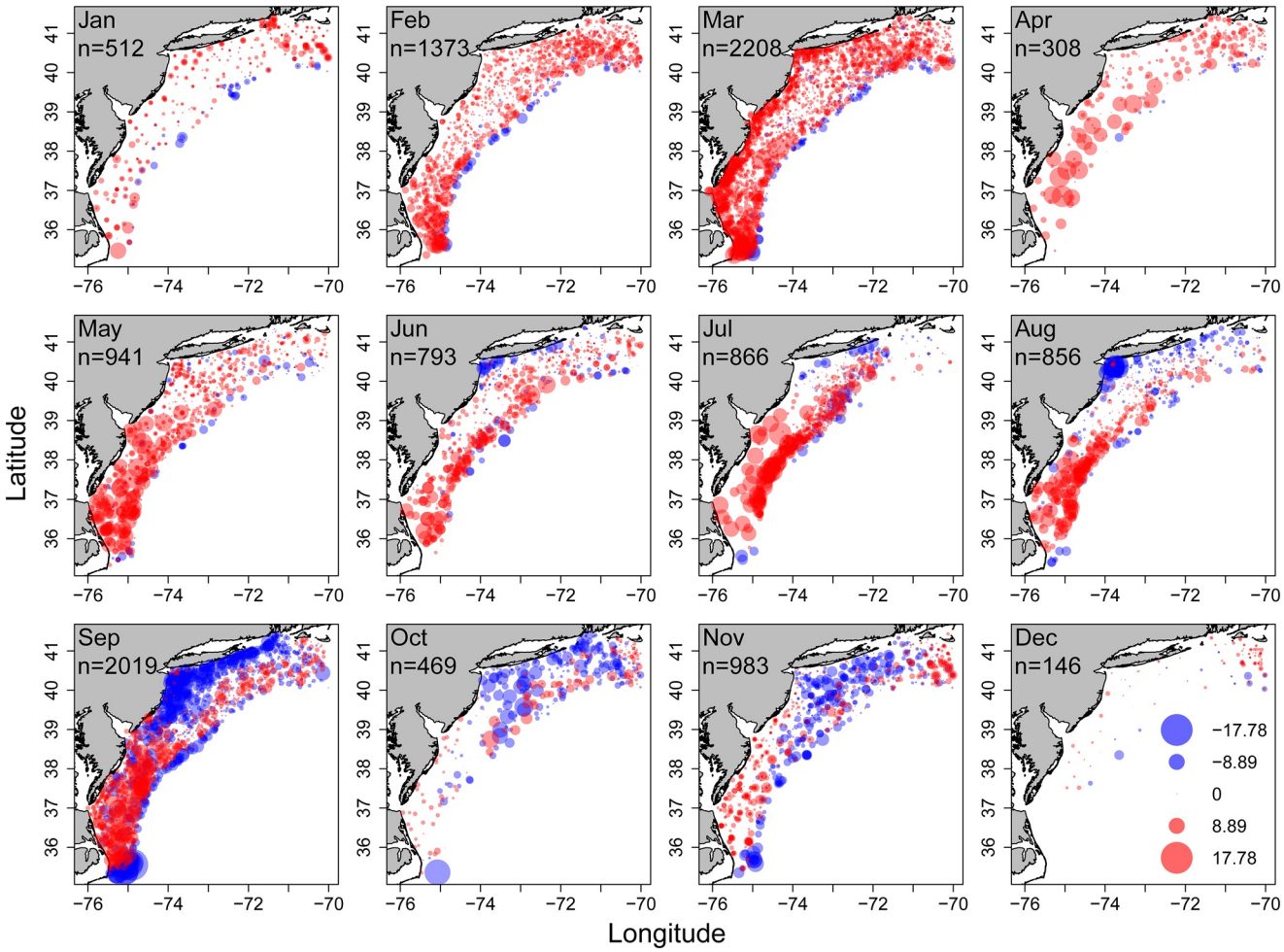

Figure 1. Differences (ROMS − CTD) between ROMS simulated bottom temperature (°C) and CTD observations at each location where a CTD observation is available in the Mid-Atlantic Bight by month for years 1980–2015. *n*: number of data points by month.

Seibert, 2013). While bias corrections are prevalent for model-simulated precipitation and surface air temperature (e.g., Teutschbein & Seibert, 2013, Table 4), bias correction of simulated ocean variables such as bottom temperature and salinity has received far less attention, even though they are frequently used in ecological studies such as habitat suitability and species distribution modeling (e.g., Chang et al., 2010; Lowen et al., 2019; Tanaka et al., 2019).

Figures 1 and 2 show the biases of daily bottom temperature and salinity simulated using the Regional Ocean Modeling System (ROMS; Kang & Curchitser, 2013, 2015) aggregated by month from 1980 to 2015 in the Mid-Atlantic Bight, off the northeast coast of the United States, from Cape Hatteras (35°N) to Cape Cod (41.5°N). These biases can be strong and exhibit spatial and temporal autocorrelations. ROMS overestimated bottom temperatures in most of the Mid-Atlantic Bight shelf from January to May, but underestimated them in the northern inshore areas from June to November, and in the deep offshore areas throughout the year. For bottom salinity, ROMS underestimated the northern and overestimated the southern portions of the Mid-Atlantic Bight. The overestimation in the south was more intense from August to November.

Although ROMS captured the seasonal patterns reasonably well, it tended to overestimate bottom temperatures in the cooler months while underestimating them when bottom temperatures were warmer (Figure 3). The mean biases (MBs) can be as large as 2.1°C for bottom temperature and −0.4ppt for bottom salinity. Spatial and temporal biases have also been documented for bottom temperatures simulated using the Finite-Volume Community Ocean Model (FVCOM) in the Northwest Atlantic Shelf region (Li et al., 2017).

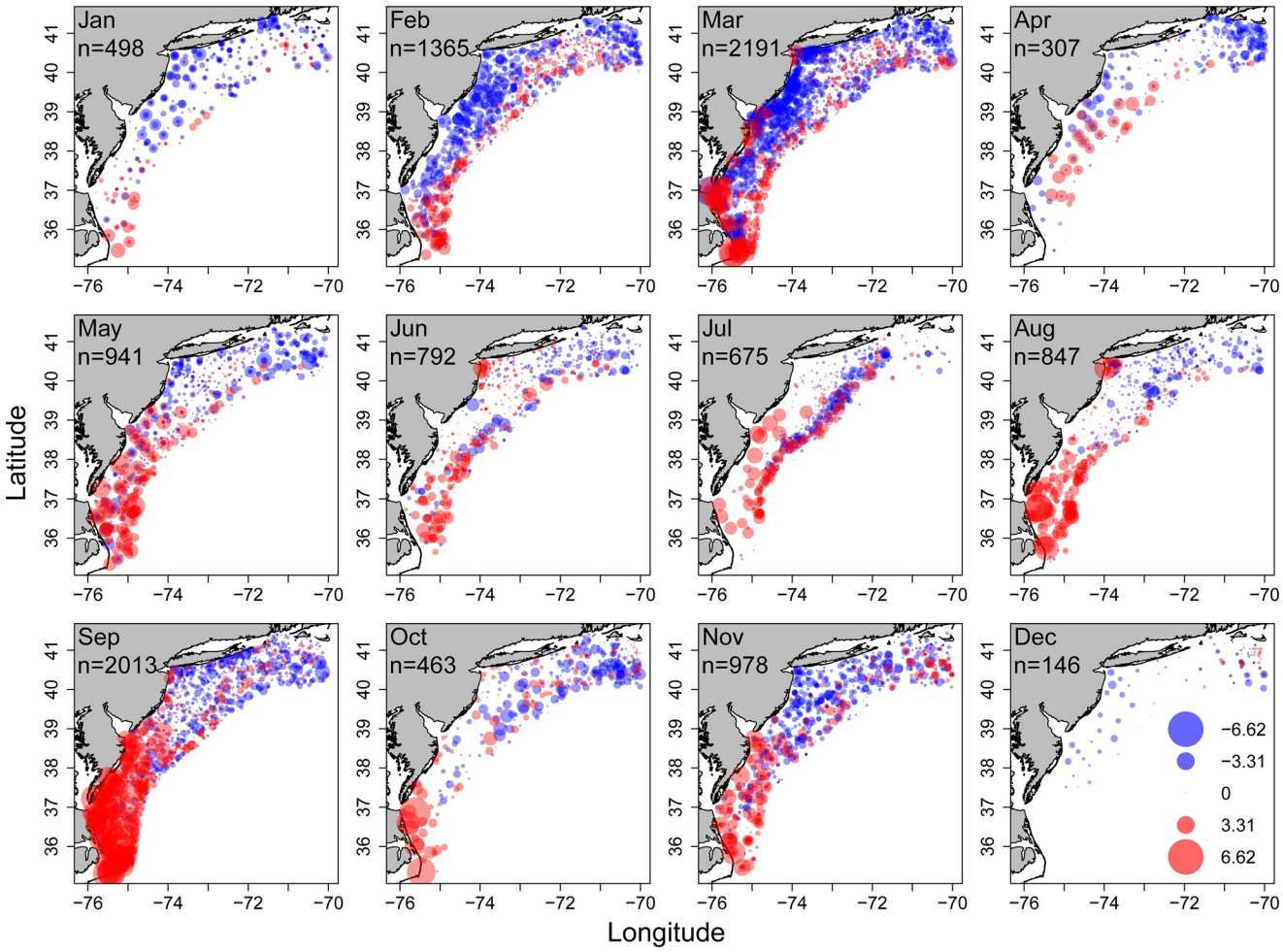

Figure 2. Differences (ROMS − CTD) between ROMS simulated bottom salinity (ppt) and CTD observations at each location where a CTD observation is available in the Mid-Atlantic Bight by month for years 1980–2015. *n*: number of data points by month.

Several methods for bias correction of climate and circulation model simulation output have been developed in the fields of climatology, meteorology, and hydrology (Lazoglou et al., 2020). Some commonly used methods include various versions of scaling approaches, quantile mapping (QM), and copulas techniques. Scaling approaches and QM are based on statistical transformations to minimize the distributional differences between the simulated and observed data (Gudmundsson et al., 2012; Teutschbein & Seibert, 2012), whereas copulas are based on modeling the complex nonlinear correlation structure between variables (Mao et al., 2015).

The performance of these established methods has been assessed by many studies (e.g., J. Chen et al., 2013; Gudmundsson et al., 2012; Gutiérrez et al., 2019; Lafon et al., 2013; Mao et al., 2015; Maraun et al., 2019; Mendez et al., 2020; Shrestha et al., 2017; Teutschbein & Seibert, 2012). Some have argued that simple methods such as delta method and linear scaling performed comparably to the more sophisticated QM (Mendez et al., 2020; Shrestha et al., 2017), while others showed that the higher-skill bias correction methods outperformed the simpler ones. All of these methods are capable of reducing the mean systematic model biases to a certain degree; however, they are much less efficient in reducing spatial-temporal patterns in the biases and thus fail to correct the spatial-temporal variability of the model simulation output (Maraun et al., 2019). None of the above methods explicitly adjust for the spatial-temporal aspects of the biases (Maraun, 2013; Maraun et al., 2019; Sunyer et al., 2015), despite the high degree of spatial-temporal autocorrelation of these biases, e.g., see Gudmundsson et al. (2012), Figure 2, Mao et al. (2015), Figures 10–12, and Johnson and Sharma (2012), Figure 1. Model simulation output with spatially and temporally autocorrelated

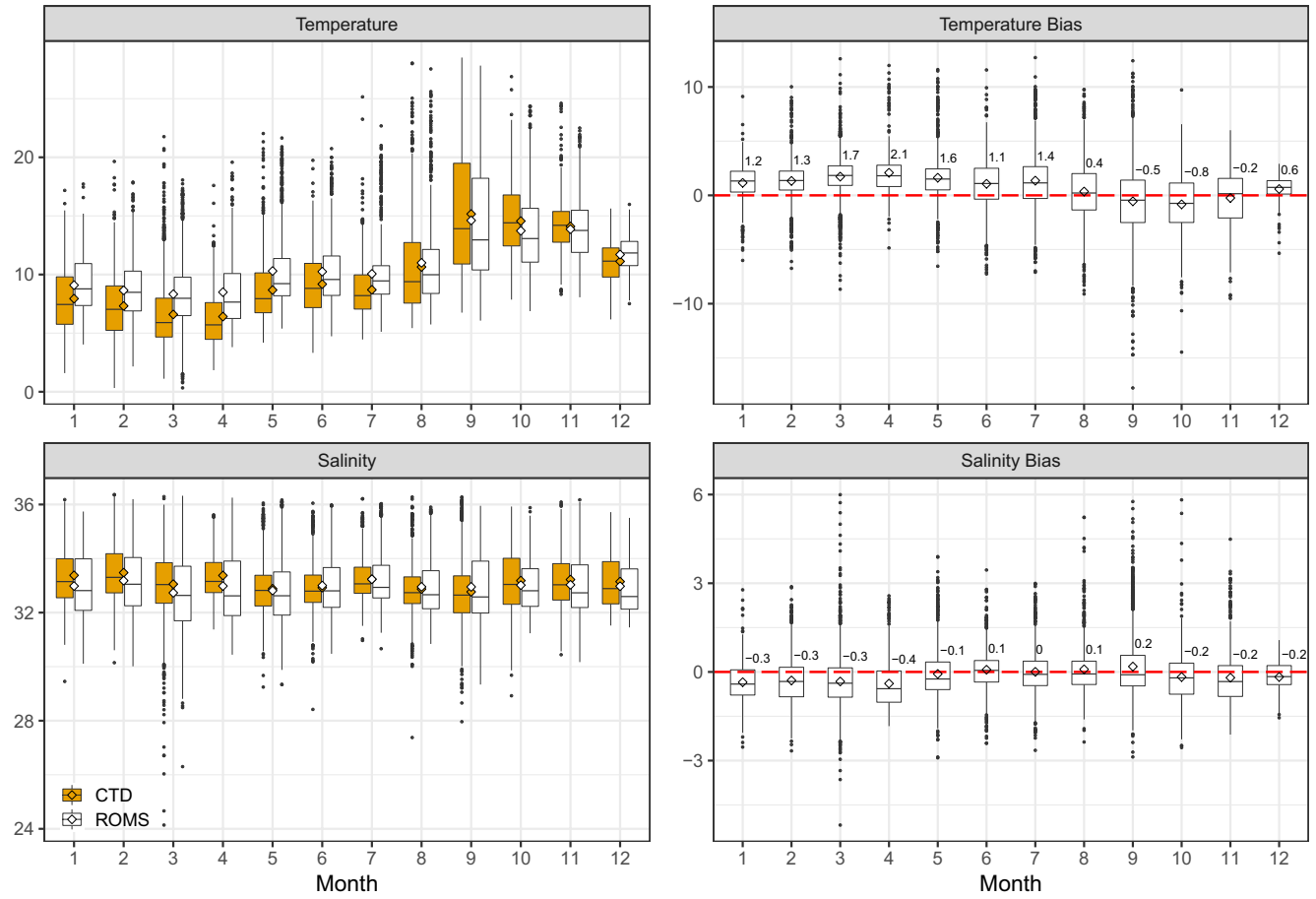

Figure 3. Boxplot and its means (diamonds) of observed CTD and ROMS simulated bottom temperature (°C) and salinity (ppt) and their differences (i.e., Bias = ROMS − CTD) in the Mid-Atlantic Bight by month for years 1980–2015. Black dots are data outside 1.5 times the inter-quartile range.

biases and underrepresented spatial-temporal variability may especially be an issue for analyses that utilize the spatial-temporal patterns of the simulation output, such as habitat suitability and species distribution modeling.

The purpose of this study is to propose using the geostatistical method of regression kriging (RK; equivalent to universal kriging or kriging with external drift) for bias correction. RK is a spatial interpolation technique that models non-stationary trends using (generalized) regression, and then takes into account the spatial dependencies of the regression residuals using ordinary kriging (OK). RK has been used in a variety of disciplines and has proven effective in terms of modeling non-stationary and spatially autocorrelated objects (Chang et al., 2017; Hengl, 2009; Webster & Oliver, 2007). Unlike many bias correction methods that directly estimate the relationships between model-simulated output and observations, we use RK to estimate the "bias" across space and time. We then use these spatial-temporal bias estimates to correct the original model simulation output. We expect that if RK can properly estimate the spatial-temporal biases, after the RK bias correction, the errors between model simulation output and observations should be reduced, and no longer show systematic spatial or temporal patterns.

Most of the commonly used criteria for evaluating bias correction techniques are focused on reducing root mean square error (RMSE) or mean absolute bias (MAB), matching the statistical moments, or maximizing the correlations between observed and simulated variables (Gudmundsson et al., 2012). However, these criteria can be problematic because they do not take into account the spatial structure of the errors (Wernli et al., 2008). For example, a bias correction that resulted in strong spatial patterns in errors would be evaluated as equal to one that lacks such patterns but has similar overall mean error levels. Ideally, the errors of

the bias-corrected model simulation output should be spatially random. Thus, the model evaluation criteria should include a measure of the extent of spatial randomness of the errors.

Several methods have been developed to evaluate structural differences in spatial characteristics such as SAL, a complex validation diagnostic that considers the structure, amplitude, and location (Wernli et al., 2008), VALUE, a comprehensive validation framework that examines many indices including spatial characteristics such as decorrelation length and variogram range (Maraun et al., 2015), and structure scores of wavelet transformed fields to capture the characteristics of field's spatial structures (Buschow et al., 2019). These methods use one or several scores to evaluate the structural differences, while other sources of errors are investigated separately. This may make it difficult to evaluate and rate the overall performance when a trade-off between reducing the overall errors and increasing the spatial similarities occurs.

Here, we introduce a simple yet informative metric, the Structural Similarity (SSIM) index, for evaluating the performance of bias correction methods. This index was originally developed to assess image quality by quantifying the differences between signals from distorted and reference images, and can simultaneously consider accuracy, precision, and spatial similarities (Wang et al., 2004). It is appropriate for evaluation of bias correction methods because the fields we examined here are similar to the natural images where pixels are highly structured and exhibit strong spatial dependencies (Wang et al., 2004). The SSIM index is used in this study to compare the accuracy, precision, and spatial similarities of the observations and bias-corrected model simulation output, which is analogous to comparing two different image signals.

In this study, we compared the performance of RK to two other bias-correcting methods: OK, which is based on an assumption of spatial stationarity, and also the most popular QM method. These bias correction methods are used to correct the biases of the ROMS simulated bottom temperature and salinity we presented in Figures 1 and 2. The performance of these bias correction methods were evaluated using SSIM as well as other standard measures such as RMSE and MAB.

The remainder of the paper is organized as follows: Section 2 introduces the example observations and model simulation output that were used for bias corrections, the theory and configurations of the RK bias correction models, and the SSIM index for model evaluations. Section 3 presents the evaluation of RK models using SSIM index along with other evaluation criteria, results of applying the RK bias corrections to model simulation output, and comparison of performances of RK models to the QM method. Section 4 discusses the advantages, limitations, and possible extensions of the RK methods and SSIM index for bias corrections; and the conclusions are summarized in Section 5.

2. Materials and Methods

2.1. Observed and Simulated Data

Hindcasts from the regional circulation model ROMS in the northwest Atlantic were used as the example for the bias-correction methods (Z. Chen & Curchitser, 2020; Z. Chen et al., 2018; Kang & Curchitser, 2013, 2015). The simulation output includes daily temperature, salinity, and other oceanographic variables from the Gulf of Mexico to the Gulf of St. Lawrence for years 1980–2015. The model grid has a 7 km horizontal spacing and 40 vertical terrains from the ocean surface to the bottom. Further details of model settings including the initial and oceanic boundary forcing, surface forcing, vertical mixing scheme, river discharges, tides, etc., are described in Kang and Curchitser (2013, 2015) and Z. Chen et al. (2018). This study focused on the bottom temperature and salinity in the Mid-Atlantic Bight (latitude: 35°N-41.5°N; longitude: 70°W-76°W; depth: 10–160 m).

We compared the ROMS simulated daily bottom temperature and salinity in the Mid-Atlantic Bight to the nearest bottom temperature and salinity observation from CTD casts in the NOAA/NEFSC Oceanography Branch Hydrographic Database. Only the CTD records that were within the ROMS prediction grid and where the depth of the CTD observation was within 10% or ±5 m of the bottom depth at the corresponding ROMS prediction point were used for comparison to the ROMS simulations. A total of 11,474 bottom temperature and 11,216 bottom salinity CTD casts met these criteria and were used in this study. The discrepancies between the observed CTD and ROMS simulated bottom temperature and salinity were assumed to be due to biases from the ROMS simulations.

2.2. Kriging Methods

Because the discrepancies between the observed CTD and ROMS output are spatially autocorrelated, we used the spatial interpolation technique of kriging to bias-correct the ROMS simulation output. Three kriging methods were evaluated: OK and two RK methods, Generalized Additive Models (GAMs; Hastie & Tibshirani, 1990; Wood, 2017) with kriged model residuals, and Generalized Additive Mixed Models (GAMMs; Zuur et al., 2009) with kriged model residuals. OK is the standard version of kriging with the following spatially stationarity assumptions (Cressie, 1986; Matheron, 1962). Let $z(t)$ be the observed value (one realization) of a stochastic process (random variable) $Z(t)$ over a domain D in $\mathbf{R}^2$. The intrinsic hypothesis for OK is:

$$E[Z(t+h) - Z(t)] = 0, \tag{1}$$

$$Var[Z(t+h) - Z(t)] = 2\gamma(h) \quad t, t+h \in D, \tag{2}$$

where t is index over space, h is distance, and $2\gamma(h) = 2[C(0) - C(h)]$ is the variogram with $C(h) = \text{cov}(Z(t+h), Z(t))$ (Matheron, 1962). Equation 1 expresses the assumption that the mean (E) is constant regardless of the location in D (spatial stationarity), whereas Equation 2 implies that the variance and covariance depends solely on the relative position (distance) of the variables $Z(t+h)$ and $Z(t)$ (Cressie, 1986).

RK extends OK to account for non-stationary trends in the mean over D (Hengl, 2009; Webster & Oliver, 2007):

$$E[Z(t)] = a + \sum_{i=1}^{k} b_i Y_i(t), \tag{3}$$

where a and b are unknown constants and k is the number of external variables $Y_i(t)$. The non-stationary mean is estimated using (generalized) regression models with external variables, after which OK is performed on the model residuals to estimate spatially autocorrelated variability (Hengl, 2009; Odeh et al., 1995). Here, we used GAM(M)s (with Gaussian distributions) as the RK regression models. GAMs extend the assumption underlying generalized linear models, that the relationship between the mean of the response variable and covariates is linear, to allow for nonlinear relationships estimated using smoothed function(s) of the predictors (Hastie & Tibshirani, 1990; Wood, 2017):

$$E[Z(t)] = a + \sum_{i=1}^{k} s_i[Y_i(t)], \tag{4}$$

where the b_i in Equation 3 is replaced by non-parametric smoothed functions s_i estimated from the data. Because of the flexible nature of this curve, GAMs can deal with highly nonlinear relationships between the response variable and covariates, and the shape of these relationships can be determined by data instead of the researcher's preconceptions (Guisan et al., 2002). However, GAMs are based on an assumption that model residuals are independent, which is unlikely to be true for our data. GAMMs use random effects to account for such autocorrelated errors in model predictions (Zuur et al., 2009). Our implementation of GAMMs employed random effects for data within a 0.5° spatial grid to account for small-scale spatial autocorrelation:

$$Z(g,t) = a + \sum_{i=1}^{k} s_i[Y_i(g,t)] + \nu_g + \epsilon_{g,t}, \tag{5}$$

where g is spatial grid index, ν_g is the random effect at spatial grid g, and $\epsilon_{g,t}$ is the individual-specific random error. Both ν_g and $\epsilon_{g,t}$ follow normal distributions with mean zero. The size of the spatial grid for the random effect was determined by selecting a size that is large enough so that there are sufficient data in the grid cells and small enough so that the data within each grid cell are similar. We ran sensitivity analysis by testing various grid sizes (0.25°, 0.75°, and 1°) for the GAMM random effect scale and compared their performances.

2.3. Model Configurations

We evaluated combinations of various response and explanatory variables to construct the regression models. Raw bias (ROMS − CTD) and relative bias ($\frac{ROMS - CTD}{CTD}$) were tested as response variables. Explanatory variables included year, month, and different combinations of depth, latitude, distance offshore, and ROMS simulated bottom temperature or salinity (Table 1). Year and month were used to account for the temporal autocorrelation of the biases within the modeling period. Depth and distance offshore were used to delineate potential east-west effect on the biases, whereas latitude was used for the north-south effect. We included ROMS simulated bottom temperature or salinity as an explanatory variable for constraining the bias correction model so that it is not completely free of the original spatial and temporal patterns simulated by the ROMS, and to account for biases being correlated with their original values. Depth and distance offshore variables were never in the same regression model because they were both used to denote the east-west effect and are strongly autocorrelated (Table 1). Depth, distance offshore, and ROMS simulated variables were evaluated separately and not included in the same regression model as single terms to avoid strong correlations among the independent variables (Table 1).

For OK applied to the biases or model residuals, we tested four candidate variogram models: spherical, exponential, Gaussian, and Matérn, and selected the one with the smallest residual sum of squares (Hiemstra et al., 2009). Before performing OK, we also checked if the variability of the data is directionally dependent (anisotropic). If so, the coordinates of the anisotropic data were rotated and rescaled to a new coordinate system so that the data become statistically isotropic (E. Pebesma et al., 2011).

To capture the temporal differences in the biases, we built the regression and kriging models for the entire time series, partitioned either seasonally, bimonthly, or monthly, and evaluated the performances of these partition lengths. The monthly models are for OKs only because our data did not cover every year for each month, and therefore they are not enough to build monthly regressions. In summary, we examined three kriging methods, different combinations of response and explanatory variables, and various temporal periods for modeling; as a result, more than 800 model combinations were tested for bias correction. The above analyses were implemented in R statistical software (R Core Team, 2020) with libraries automap (E. J. Pebesma, 2004), intamap (Hiemstra et al., 2009), gstat (E. Pebesma et al., 2011), and mgcv (Wood, 2017).

2.4. QM Method

The QM we implemented here is described in Gudmundsson et al. (2012) as the nonparametric smoothing splines transformation approach, which has the highest skill to reduce systematic biases in their study. The observed CTD and ROMS output were aggregated into 100 quantiles to avoid overfitting. We then fit a smooth spline to the quantile-quantile relationship between the observed and simulated data and used the estimated spline smoother to adjust the distribution of the ROMS output to match the distribution of the CTD observations (Gudmundsson, 2016). The QM was performed monthly for the bottom temperature and salinity because of the differences in their CDFs by month. The QM analysis was implemented in R statistical software (R Core Team, 2020) using the qmap library (Gudmundsson, 2016).

2.5. Evaluation of Model Performance

We evaluated the performances of the RK methodologies primarily by using RMSE, the *p*-value of Moran's *I* statistic (MI), and the SSIM index, while also providing MAB, MB, and mean relative bias (MRB). RMSE is calculated as the square root of the mean of the squared residuals (i.e., differences) between observed CTD and (bias-corrected) ROMS output. It is one of the most widely used metrics for accuracy and precision, but it does not measure whether there are spatial patterns in the residuals. MI is used to measure the degree of spatial autocorrelations in the residuals. Low MI values indicated high spatial autocorrelations. However, it does not measure overall accuracy or precision. The SSIM index is used in this study to concurrently examine the accuracy, precision, and spatial similarities of the (bias-corrected) ROMS output as compared to the CTD observations. The SSIM index separately but simultaneously compares the differences between the "image" signals' mean, variance, and correlations to represent "luminescence" (l), "contrast" (c), and "structural information" (s), respectively:

Table 1
List of Combinations of Various Explanatory Variables Tested in the Regressions

Combination	Explanatory variables
1	Year; Month; Dist
2	Year; Month; Depth
3	Year; Month; ROMS
4	Year; Month; Lat
5	Year; Month; Dist; Lat
6	Year; Month; Depth; Lat
7	Year; Month; ROMS; Lat
8	Year; Month; (Dist/Lat)
9	Year; Month; (Depth/Lat)
10	Year; Month; (ROMS/Lat)
11	Year; Month; (ROMS/Depth)
12	Year; Month; (ROMS/Dist)
13	Year; Month; Dist; (ROMS/Lat)
14	Year; Month; Depth; (ROMS/Lat)
15	Year; Month; ROMS; (Dist/Lat)
16	Year; Month; ROMS; (Depth/Lat)
17	Year; Month; Lat; (ROMS/Dist)
18	Year; Month; Lat; (ROMS/Depth)

Note. Dist, distance offshore; ROMS, ROMS simulated bottom temperature or salinity; Lat, Latitude.

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \tag{6}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{7}$$

$$s(x,y) = \frac{2\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \tag{8}$$

where x and y are "image" signals, which for our purposes are observed CTD data and (bias-corrected) ROMS output, μ_x, μ_y, σ_x, σ_y, and σ_{xy} are their means, standard deviations, and covariance, and C_1, C_2, and C_3 are small constants to avoid instability when the denominators are very close to zero. The SSIM index is the combination of the three components:

$$S(x,y) = [l(x,y)]^\alpha [c(x,y)]^\beta [s(x,y)]^\gamma, \tag{9}$$

where α, β, and γ are weights used to adjust the relative importance of the three components. Larger SSIM indices indicate close correspondence between observations and the simulations. If the observed CTD and ROMS output are identical, the SSIM index is 1. A simplified version of SSIM was used for this study where $\alpha = \beta = \gamma = 1$ and $C_1 = C_2 = C_3 = 0$. The same weights were used for α, β, and γ so that the models selected using this SSIM have a balanced ability to produce accurate, precise, and spatially similar model estimates relative to data.

Ten-fold cross-validation was used to obtain the performance statistics such as RMSE, MI, and SSIM. Data from all years were randomly partitioned into 10 sets. One set was left out for testing and validating, while the rest were used for training the model. This procedure was iterated 10 times. Model predictions from the 10 testing sets were used to calculate the performance statistics. All the performance statistics were calculated by month for each year and then averaged to represent the overall performance of each model.

3. Results

3.1. Performance Statistics

The SSIM index was highly correlated with both RMSE and MI, whereas the relationship between RMSE and MI was less clear (Figure 4). The model with the highest MI, and hence lowest spatial autocorrelation, had a poor overall fit, that is, high RMSE, for both temperature and salinity (Figure 4). Thus, there tends to be a trade-off between reducing spatial autocorrelation of the residuals and improving precision and accuracy of the models. The maximum SSIM tends to occur where RMSE is low, albeit not at its minimum, and MI is high, but not at its maximum (Figure 4), so it reflects both good precision and accuracy combined with low spatial autocorrelation.

3.2. Comparison of Kriging Methods and Model Configurations

The GAMM RK model had the highest SSIM score for both temperature and salinity with raw bias as the response variable, and year, month, depth, and ROMS simulated variable/latitude interaction term as the explanatory variables (Combination 14; Table 1). The bimonthly GAMM (R^2: 0.41–0.67) with monthly kriged GAMM residuals performed the best for bias-correcting temperature, whereas the seasonal GAMM (R^2: 0.49–0.66) with monthly kriged GAMM residuals were the best for bias-correcting salinity.

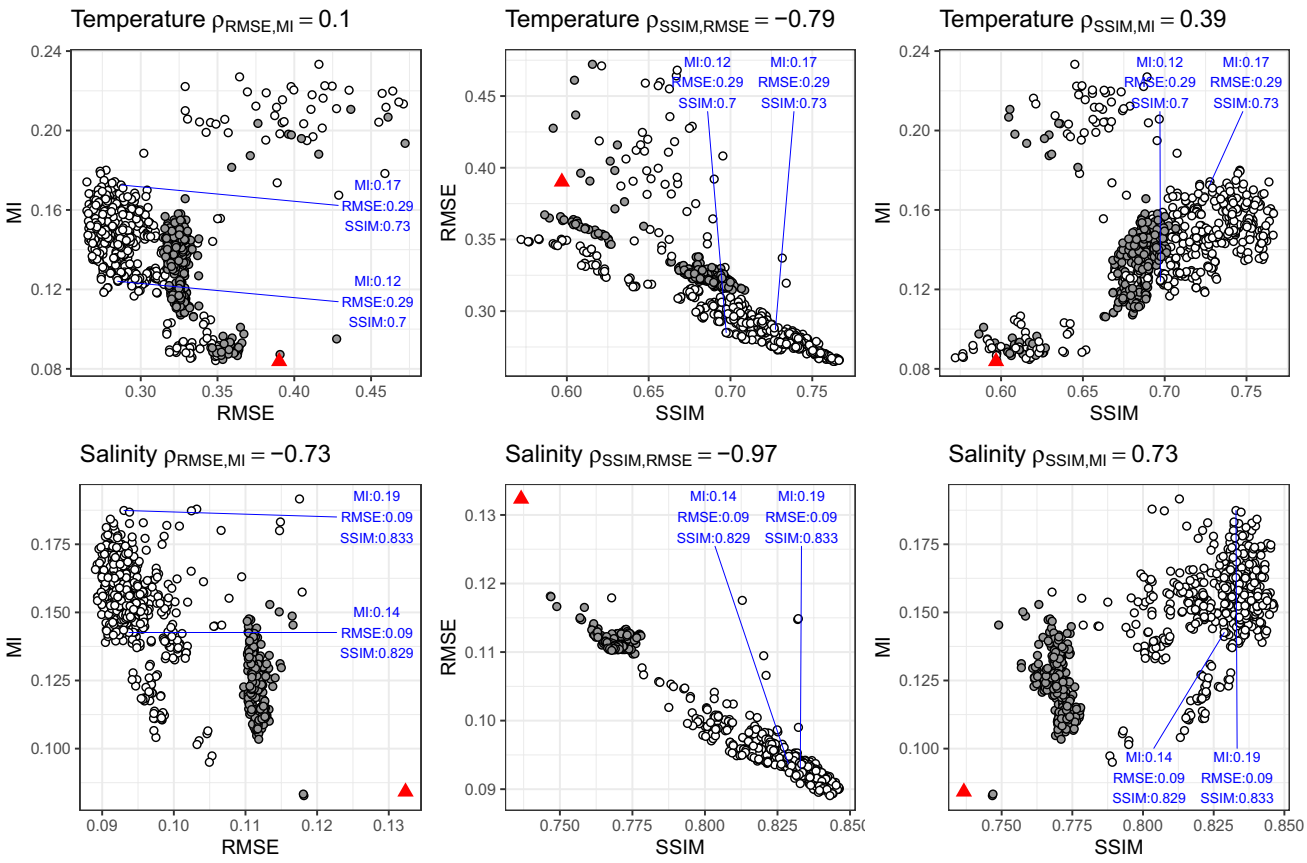

Figure 4. Comparisons and Pearson correlation coefficients (ρ) between averaged RMSE, MI, and SSIM for all candidate models. The statistics of RK models with and without ROMS simulated bottom temperature (°C) or salinity (ppt) as a variable in the regression are circles in white and gray, respectively, whereas the QM models are triangles in red.

Among the kriging methods we tested, both the GAM and GAMM RK methods significantly outperformed the OK and the GAMM RK typically were slightly better than the GAM RK (Figure 5). Model configuration had strong impacts on the performance of bias correction models for temperature, but less so for salinity (Figure 5). Using raw bias as a response variable was better than relative bias; however, this has more influence on temperature than salinity (Figure 5). Of all the external variables tested in the regressions, the ROMS simulated variable (bottom temperature or salinity) was the most influential (Figure 5). The models that did not include ROMS simulated variable as a predictor have significantly lower SSIMs for both temperature and salinity (Figures 4 and 5). Models with latitude or depth slightly outperformed the other models for both temperature and salinity (Figure 5). Shorter modeling periods (e.g., month) were better for bias-correcting temperature, but the choice of modeling period had an insignificant impact on the bias correction of salinity (Figure 5). The performances of GAMMs were insensitive to the spatial grid size for the random effect scale and only slightly decreased with increasing grid size (Figure 6).

3.3. ROMS Bottom Temperature and Salinity

Table 2 summarizes the performance statistics before and after bias corrections for the ROMS simulated bottom temperatures and salinities. The original uncorrected biases, when averaged over space and time, were low to moderate. MAB, MB, and MRB were 2.07°C, 0.71°C, and 16% for temperature and 0.66 ppt, −0.12 ppt, and 0.004% for salinity, respectively (Table 2). However, the biases for both temperature and salinity were strongly autocorrelated in space and time (Figures 1 and 2), as reflected by MI values that were close to zero for both temperature (0.1) and salinity (0.08; Table 2).

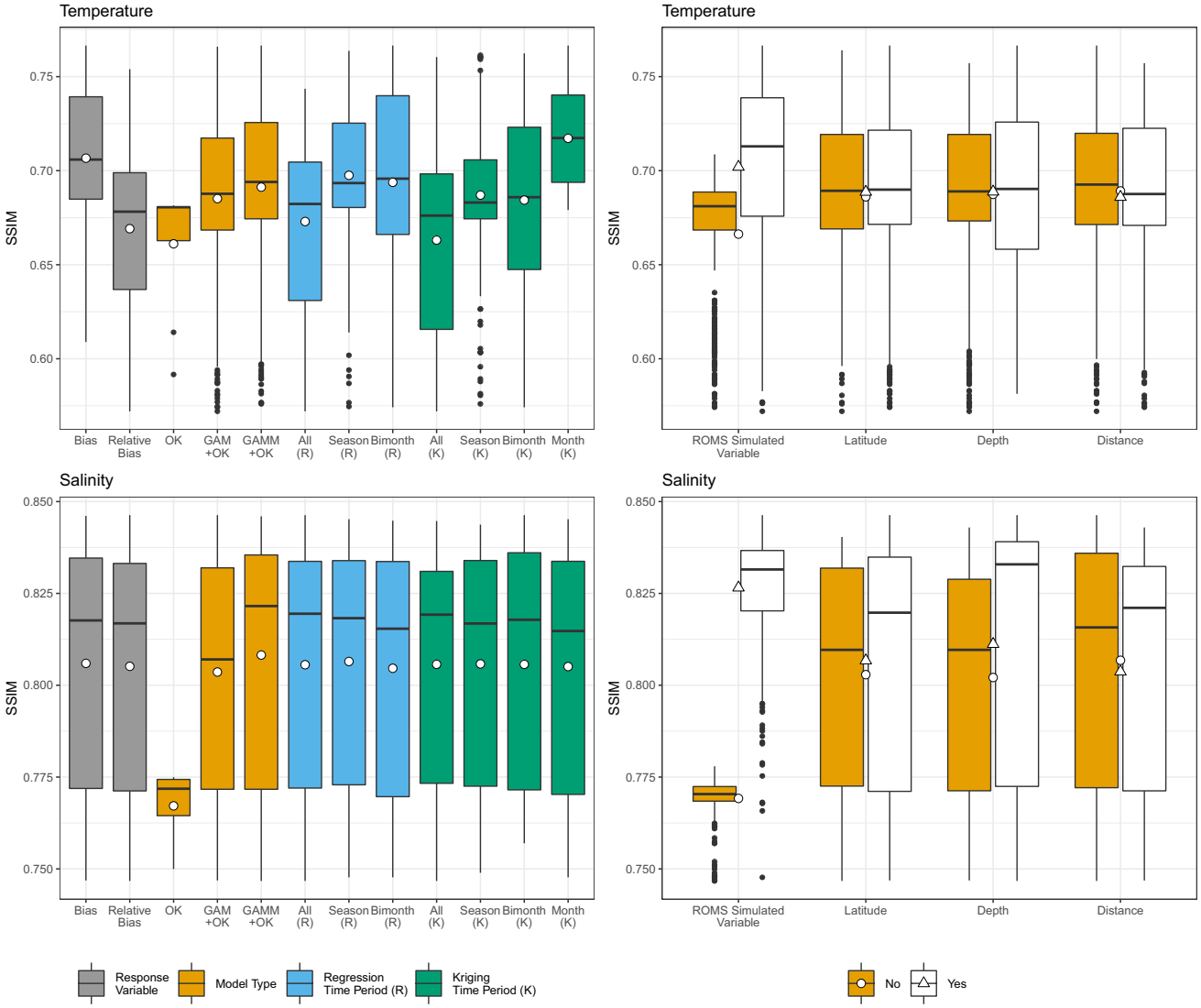

Figure 5. Boxplots and its means (open circles or triangles) of SSIM by model configurations (left panel) and RK regressions with (Yes) or without (No) variables including ROMS simulated variables, latitude, depth, and distance offshore (right panel) for bottom temperature (°C) and salinity (ppt). Black dots are data outside 1.5 times the inter-quartile range.

The GAMM RK model significantly improved the precision and accuracy of the ROMS simulations by reducing the MAB, MB, MRB, and RMSE by 44%, 100%, 83%, and 39% for temperature and 37%, 110%, 112%, and 35% for salinity, respectively (Table 2). MB and MRB were close to zero for both temperature and salinity after the bias correction (Table 2). Mean and variation of the ROMS simulated bottom temperature and salinity after the GAMM RK bias correction were very close to the CTD observations (Table 2).

The GAMM RK method reduced spatial autocorrelation of the biases, with MI increasing by 68% and 146% for temperature and salinity, respectively (Table 2). It successfully removed the patchiness and structure of the biases, along with reducing their magnitude, as reflected by the improvement of the SSIM values for both temperature (29%) and salinity (14%; Table 2 and Figures 7 and 8). The temporal biases were also largely reduced, while the temporal temperature and salinity trend were maintained (Figures 9 and 10). However, the bias corrections were less effective around the inshore and offshore edge (shelf break) of the study area (Figures 7 and 8). We suspect that this is due to fewer CTD observations available in those areas.

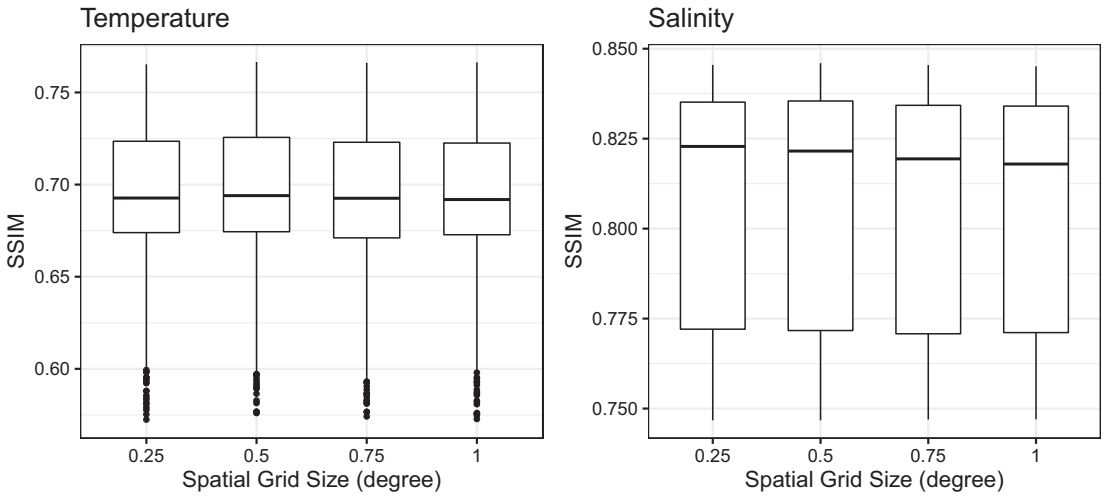

Figure 6. Boxplots of SSIM by spatial grid size for the GAMM random effect scale for bottom temperature (°C) and salinity (ppt). Black dots are data outside 1.5 times the inter-quartile range.

Figures 11 and 12 show examples of averaged ROMS daily bottom temperature and salinity before and after the GAMM RK bias correction in the Mid-Atlantic Bight in March and September of 2015. The bias corrections increased the contrast of the temperature surface and reduced the contrast of the salinity surface, while preserving the original patterns of the surfaces (Figures 11 and 12). Comparing the biases in Figures 1, 7 and 11, the GAMM RK model captured the over/underestimation of temperature in March and September well. The bias correction of salinity was not as successful as temperature. Although it effectively

Table 2
Summary of Mean (SE, Minimum-Maximum) of the Monthly Observed CTD Means, Original, and Bias-Corrected Means, Performance Statistics, and Percent Change of the Performance Statistics' Mean (SE) of ROMS Simulated Bottom Temperature (°C) and Salinity (ppt) for Years 1980–2015

			Bias-corrected (*BC*)		$(BC - O)/O \times 100\%$	
Variable	Statistics	Original (*O*)	GAMM + OK	QM	GAMM + OK	QM
Temperature	Mean (CTD)	10.15 (0.22, 6.73–13.10)	–	–	–	–
	Mean (ROMS)	10.86 (0.17, 8.86–13.44)	10.15 (0.19, 7.29–12.25)	10.15 (0.20, 7.63–13.05)	−6.59 (7.54)	−6.58 (14.50)
	MAB	2.07 (0.09, 1.43–4.46)	1.16 (0.03, 0.87–1.55)	1.71 (0.08, 1.16–3.90)	−43.74 (−63.05)	−17.47 (−5.75)
	MB	0.71 (0.14, −0.97 to 3.52)	−0.00 (0.08, −0.85 to 0.91)	−0.00 (0.13, −1.62 to 2.52)	−100.27 (−45.17)	−100.00 (−8.91)
	MRB	0.16 (0.02, 0.01–0.53)	0.03 (0.01, −0.07 to 0.14)	0.03 (0.01, −0.12 to 0.31)	−82.76 (−54.07)	−81.49 (−24.83)
	RMSE	0.43 (0.02, 0.27–1.14)	0.27 (0.01, 0.19–0.42)	0.38 (0.02, 0.26–1.02)	−38.62 (−57.50)	−12.24 (−5.95)
	MI	0.10 (0.02, 0.00–0.53)	0.17 (0.02, 0.01–0.56)	0.09 (0.02, 0.00–0.39)	67.62 (18.80)	−7.81 (−17.70)
	SSIM	0.59 (0.02, 0.42–0.79)	0.77 (0.01, 0.60–0.88)	0.61 (0.02, 0.43–0.80)	29.04 (−28.47)	1.90 (5.18)
Salinity	Mean (CTD)	33.12 (0.07, 32.22–33.77)	–	–	–	–
	Mean (ROMS)	33.00 (0.07, 32.27–34.72)	33.13 (0.05, 32.32–33.67)	33.13 (0.07, 32.35–34.77)	0.41 (−24.66)	0.40 (−7.37)
	MAB	0.66 (0.03, 0.39–1.37)	0.41 (0.01, 0.30–0.60)	0.58 (0.03, 0.38–1.35)	−37.02 (−55.71)	−11.30 (−2.10)
	MB	−0.12 (0.06, −0.72 to 1.21)	0.01 (0.04, −0.34 to 0.47)	0.01 (0.06, −0.57 to 1.25)	−109.31 (−42.93)	−106.86 (−2.65)
	MRB	−0.00 (0.00, −0.02 to 0.04)	0.00 (0.00, −0.01 to 0.01)	0.00 (0.00, −0.02 to 0.04)	−111.97 (−45.64)	−108.34 (−5.95)
	RMSE	0.14 (0.01, 0.09–0.34)	0.09 (0.00, 0.06–0.14)	0.13 (0.01, 0.08–0.35)	−34.70 (−56.22)	−7.97 (2.69)
	MI	0.08 (0.01, 0.00–0.24)	0.19 (0.02, 0.00–0.54)	0.08 (0.01, 0.00–0.29)	146.75 (97.98)	9.59 (−2.97)
	SSIM	0.74 (0.02, 0.47–0.89)	0.84 (0.01, 0.62–0.94)	0.74 (0.02, 0.51–0.87)	13.82 (−37.86)	0.05 (−6.57)

Note. SE of the means of the CTD data and ROMS output are proportional to the deviation/variability of these means, but it is not a measure of the reliability of the underlying data.

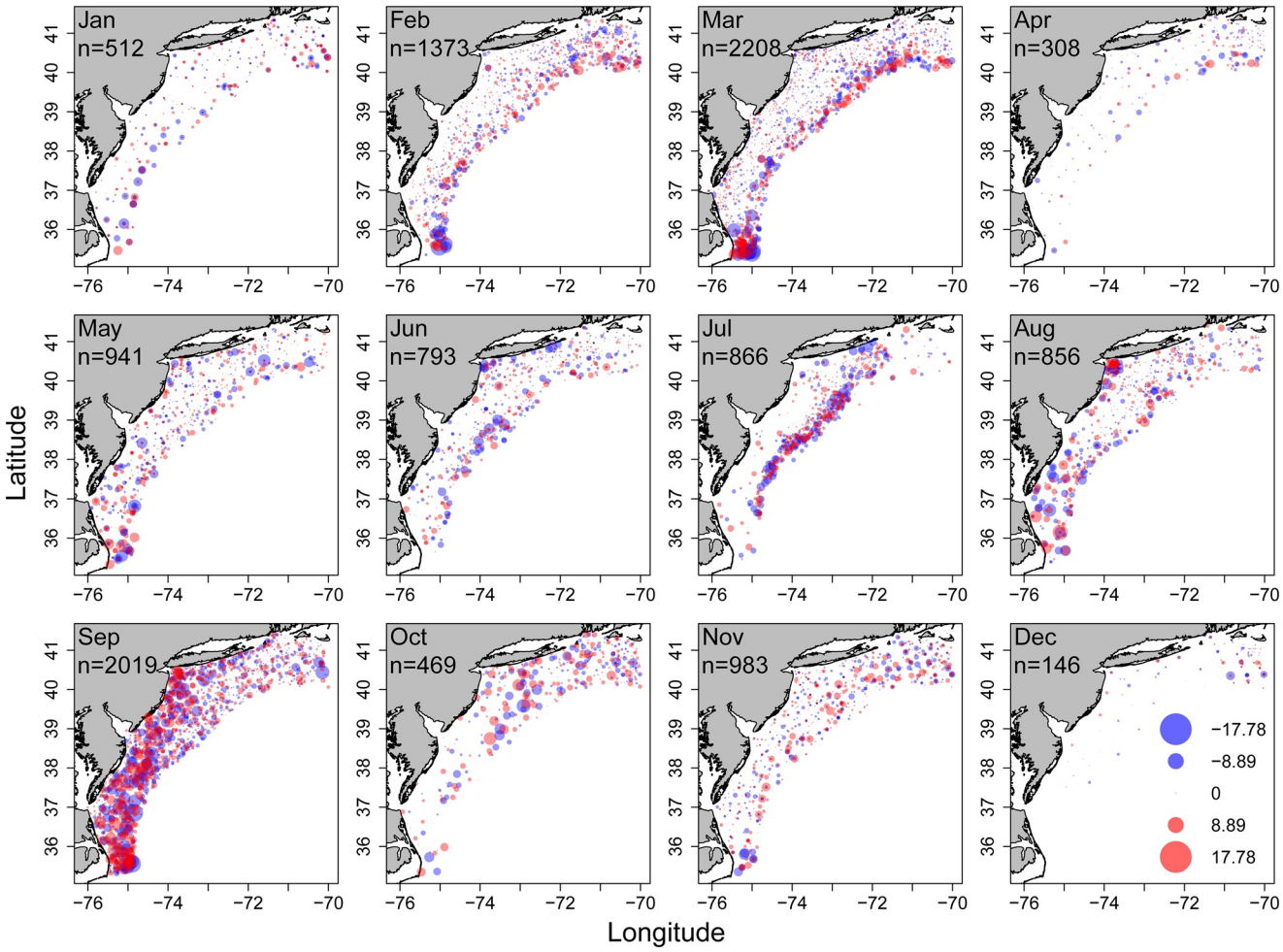

Figure 7. Differences (ROMS − CTD) between GAMM RK bias-corrected ROMS simulated bottom temperature (°C) and CTD observations at each location where a CTD observation is available in the Mid-Atlantic Bight by month for years 1980–2015. *n*: number of data points by month.

corrected for the underestimated salinity in the northern Mid-Atlantic Bight, it did not fully correct for the overestimation of salinity in the south in March (Figures 2, 8 and 12).

3.4. QM Method

The QM technique of using smooth splines to transfer the quantiles performed well for both temperature and salinity with respect to averaged bias (Table 2 and Figure 13). The QM reduced MAB, MB, and MRB by 18%, 100%, and 82% for temperature and 11%, 107%, and 108% for salinity, respectively, similar to the performance of GAMM RK models (Table 2). Moreover, like the GAMM RK models, the mean and variation of QM bias-corrected bottom temperature and salinity were very close to the observations (Table 2). However, QM was not able to reduce the spatial-temporal autocorrelation and SE of the biases as well as RK. MI was worse (reduced 8%) after the QM bias correction for temperature (Table 2). The QM only improved SSIM by 2% for temperature and 0.05% for salinity, much less than most of the RK models (Table 2 and Figure 4).

4. Discussion

In this study, we identified spatiotemporal biases in the ROMS simulated bottom temperatures and salinities for the Mid-Atlantic Bight, and corrected them using alternatively the OK, RK, and QM methods. Although both GAMM RK and QM effectively bias-corrected the overall mean and variation of temperature and salinity, only RK substantially reduced the spatial-temporal autocorrelation and SE of the biases, while

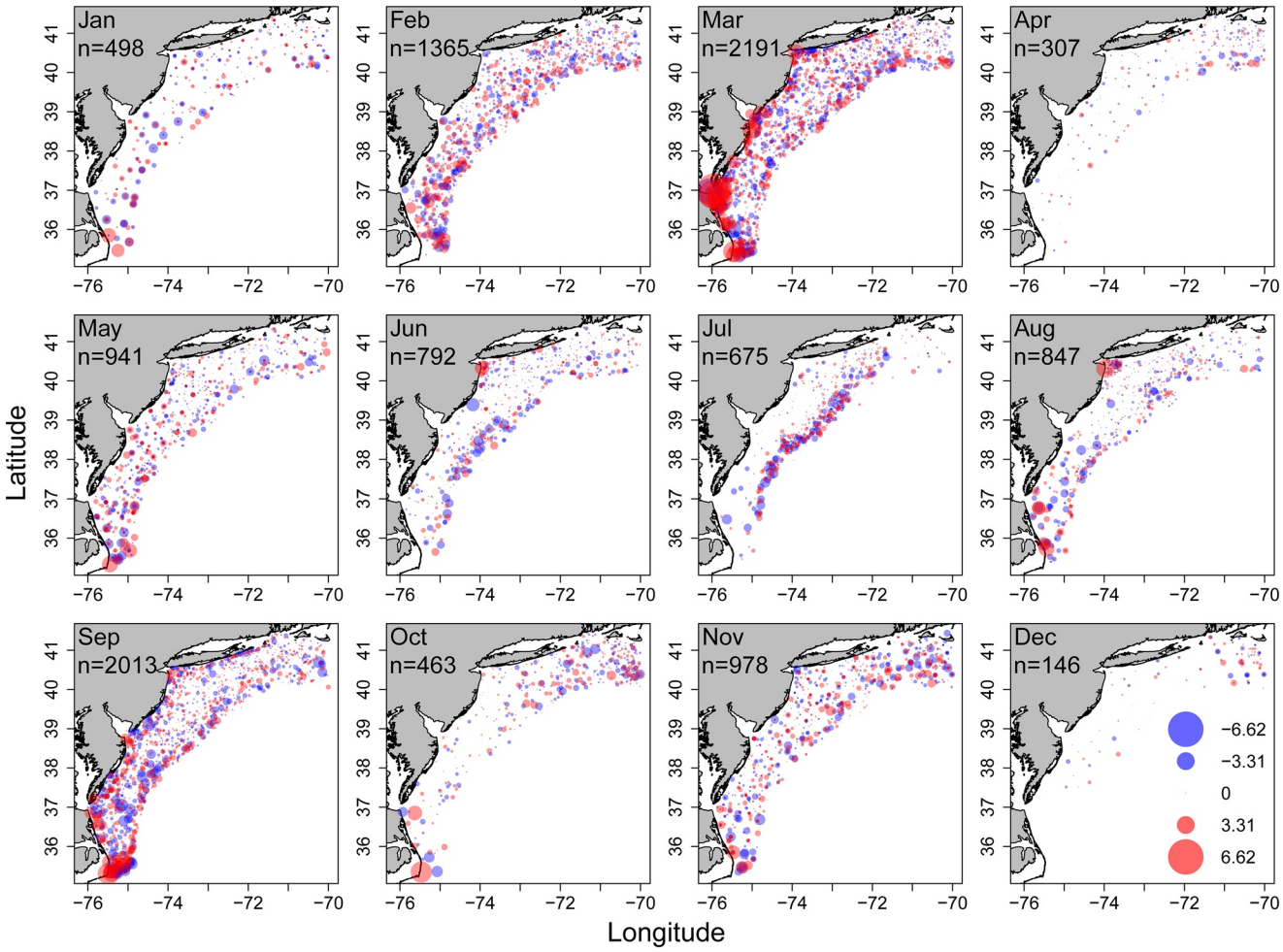

Figure 8. Differences (ROMS − CTD) between GAMM RK bias-corrected ROMS simulated bottom salinity (ppt) and CTD observations at each location where a CTD observation is available in the Mid-Atlantic Bight by month for years 1980–2015. *n*: number of data points by month.

at the same time preserving the original spatiotemporal patterns of the ROMS simulated bottom temperature and salinity surfaces.

The RK method decomposes the observed surface into large-scale trends and small-scale variations. The large-scale non-stationary trend is modeled using regressions, whereas the small-scale variation is modeled using OK on the regression residuals. Based on OK's intrinsic hypothesis (Equations 1 and 2), the small-scale variation is assumed to be stationary, with its covariance depends only on distance between data points. By their construction, it would be expected that errors in climate and circulation models will be autocorrelated; this spatial autocorrelation can be captured by the OK variogram. Therefore, RK is suitable and effective for bias correction of climate and circulation model simulation output, especially when biases are highly spatially autocorrelated with non-stationary large-scale trends, as is often the case.

However, for RK to be effective, the OK stationary assumption for the regression residuals needs to be met. In other words, the chosen regression model needs to eliminate large-scale trends so that the remaining residuals are stationary. Violating the stationarity assumption may cause the estimation of the variogram to become unstable and reduce the effectiveness and predictive power of RK. For this reason, when OK is used directly without a regression model, it is only slightly better than the QM. The mean SSIM of all the tested OKs was 0.66 for temperature and 0.77 for salinity (Figure 5), whereas the SSIM for QM was 0.61 for temperature and 0.74 for salinity (Table 2). Mean performance of OK was also similar to the RK models that performed poorly, such as the RK models without ROMS simulated variables, which have mean SSIM 0.67 for temperature and 0.76 for salinity (Figure 5).

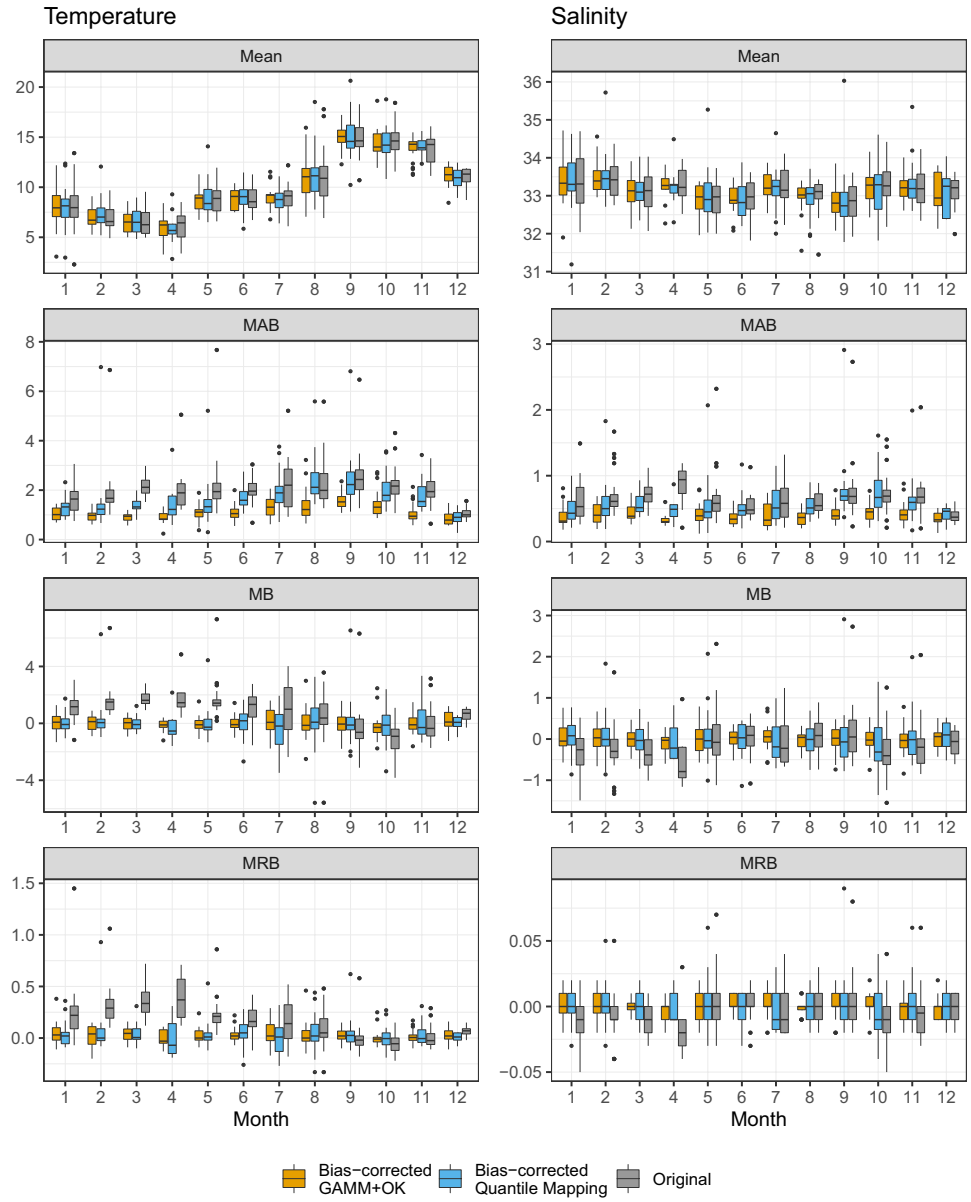

Figure 9. Boxplots of monthly mean, MAB, MB, and MRB of ROMS simulated bottom temperature (°C) and salinity (ppt) before and after bias corrections in the Mid-Atlantic Bight for years 1980–2015. Black dots are data outside 1.5 times the inter-quartile range.

In RK, the large-scale variation can be modeled using any type of regression which can be chosen based on the characteristics of the target simulated variables. For example, when the target variable is zero-inflated and over-dispersed, such as often the case for precipitation, two-stage hurdle regressions can be used (e.g., Chang et al., 2017; Zuur et al., 2009).

A major advantage of RK is its ability to incorporate auxiliary information through covariates in the regression (Teutschbein & Seibert, 2012), so variables such as depth and month can be incorporated to inform the bias corrections. Covariates can also be used to constrain output so that the bias correction does not completely eliminate the spatiotemporal patterns from the climate and circulation models, which can be a concern when using RK because it explicitly adjusts the spatial-temporal structure of the target simulated variables (Cannon, 2016; Maraun, 2016; Maraun et al., 2019). The additional information from the external variables is also useful when estimating biases at the locations where there are few nearby observations.

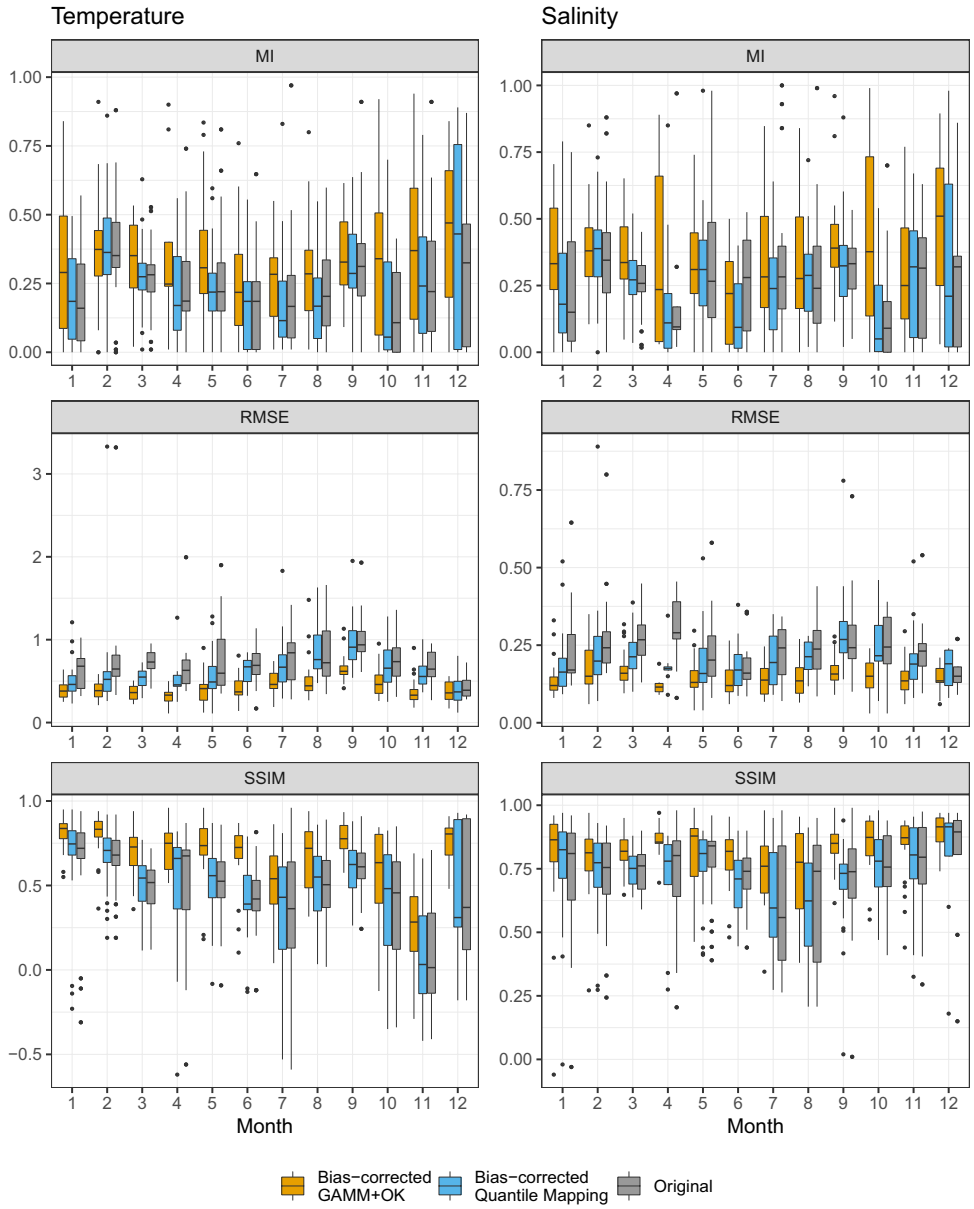

Figure 10. Boxplots of monthly MI, RMSE, and SSIM of ROMS simulated bottom temperature (°C) and salinity (ppt) before and after bias corrections in the Mid-Atlantic Bight for years 1980–2015. Black dots are data outside 1.5 times the inter-quartile range.

In the three kriging methods, we modeled differences between observed and simulated bottom temperature and salinity, and then bias-corrected by adding the model estimates to the ROMS simulation output. Although this approach is logical since our main focus was to correct systematic biases in the climate and circulation model simulation output, it is possible that this process could produce values that are outside the reasonable range of target simulated variables. The same issue may occur for the QM method as well. We included the ROMS simulated variable (bottom temperature or salinity) as an explanatory variable in the RK regressions to indirectly constrain the bias estimates through their correlations. This had a substantial impact on model performance (Figure 5); the RK corrected temperature and salinity using that covariate were within reasonable ranges (Figure 9). Including the original target simulated variables in RK bias correction models is crucial, especially when the biases are directly modeled.

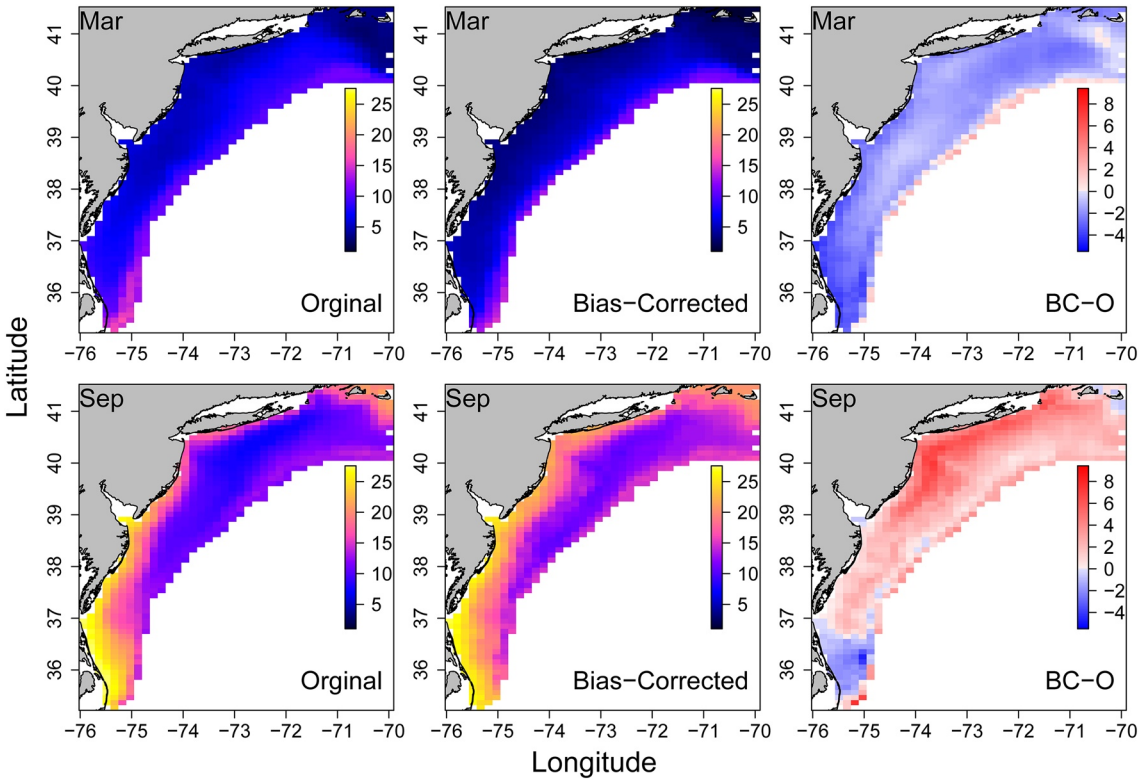

Figure 11. Averaged ROMS simulated daily bottom temperature (°C) before and after GAMM RK bias corrections and their differences (BC-O) in the Mid-Atlantic Bight in March and September of 2015.

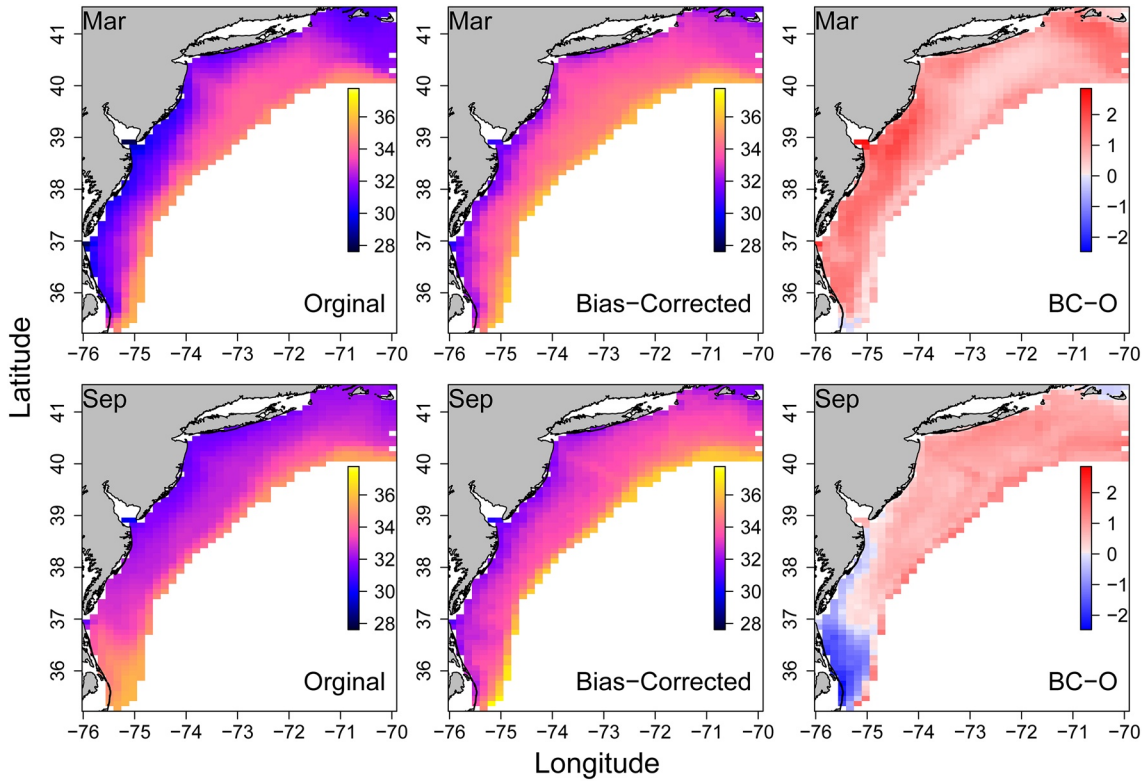

Figure 12. Averaged ROMS simulated daily bottom salinity (ppt) before and after GAMM RK bias corrections and their differences (BC-O) in the Mid-Atlantic Bight in March and September of 2015.

Spatial variability was modeled using covariates in the RK regression (explicitly as latitude or implicitly as depth or distance offshore), and in the kriging variogram for both RK and OK. Temporal variability was only accounted for in RK by using year and month as covariates in the regressions, and by estimating the regressions and OKs by temporal periods (entire time series, seasonal, bimonthly, or monthly). Although this approach eliminates much of the temporal trends in the biases, some small scale temporal variability remains (Figures 9 and 10). This potentially could be addressed using spatial-temporal kriging, which employs a variogram that is as a function of both space and time, while large-scale trends (if any) can still be estimated using regressions (Ruybal et al., 2019).

Although the RK method has many advantages, it is more complex, computationally intensive, and requires better data coverage than QM and other simple scaling approaches. The use of RK is especially important for applications such as habitat suitability or species distribution modeling that use fine-scale spatial-temporal simulation output for their analyses. On the other hand, QM and some other simple scaling methods can robustly correct the (spatially aggregated) mean and variation of the target simulated variables (e.g., Lafon et al., 2013; Mendez et al., 2020; Shrestha et al., 2017; Sunyer et al., 2015), and can be used if only these quantities are of interest.

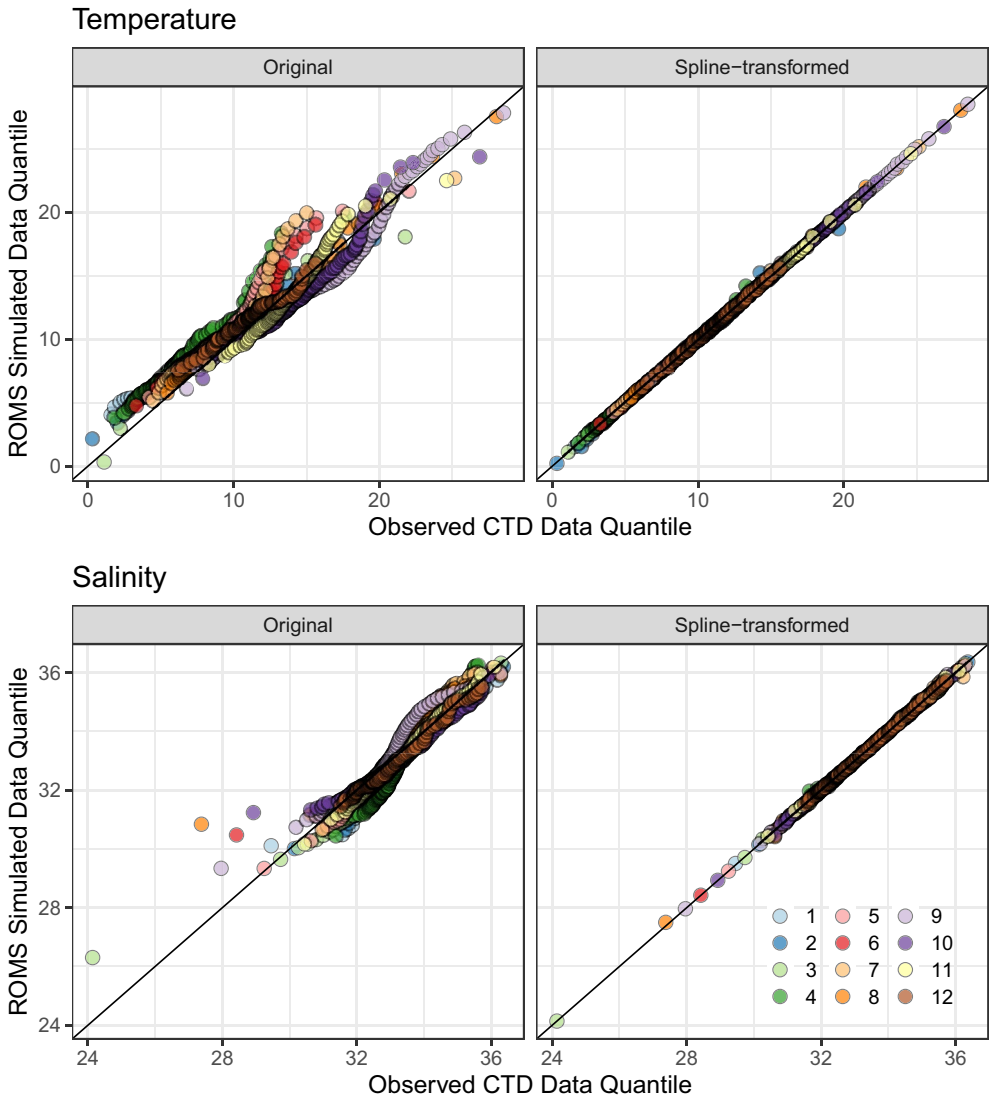

Figure 13. Quantile-quantile plots of observed CTD and ROMS simulated bottom temperature (°C) and salinity (ppt) before (left panel) and after (right panel) QM bias corrections in the Mid-Atlantic Bight by month for years 1980–2015.

While the RK method is effective in removing biases in hindcasts, it is also of interest to bias-correct the forecasts of future conditions from climate and circulation models. For forecasts, either the terminal year, or an average over the last several years, could be used in the year covariate in the regression for future years. Alternatively, it could be assumed that the temporal trends observed during the hindcast will extend into the future, although such extrapolations should be implemented with caution. These alternatives need to be evaluated carefully, and perhaps several approaches should be used to help understand a source of uncertainty.

Our results demonstrate that the SSIM index, which was originally developed for image quality assessment, can also be used for evaluation of bias correction methods, and is especially useful when spatial structure is of interest. Unlike other metrics such as RMSE and MI, SSIM can simultaneously account for the precision, accuracy, and spatial similarity of the biases. For example, models with the same RMSE but different spatial structure can be distinguished using SSIM (Figure 4). A number of variants of SSIM may also be useful, such as changing the values of α, β, and γ for highlighting one aspect over the others (Equation 9), or calculating SSIM within local windows with spatially varied weights to the mean, variance, and correlation coefficients, and then averaging the local SSIMs to get an overall SSIM measure (Wang et al., 2004).

5. Conclusions

We evaluated three types of kriging as well as a QM method for their abilities to bias-correct ocean circulation model output. These methods performed similarly in correcting the mean and variation of the ROMS bottom temperature and salinity, but showed substantial differences in reducing the spatial-temporal autocorrelation and SE of the biases. The GAMM RK method was found to be the best, and it can simultaneously reduce not only the overall mean and SE of the bias, but also its spatial-temporal autocorrelation. This method considerably improved the spatiotemporal similarity between the observed CTD and ROMS simulated bottom temperature and salinity. The RK approach is very flexible, and for that reason can be easily adapted to other climate and circulation model simulation output. This work has profound implications for studies that use the output from such a model for fine-scale mapping, for example, to identify physical characteristics such as the cold pool in the Mid-Atlantic Bight (Z. Chen & Curchitser, 2020; Z. Chen et al., 2018) or to determine habitat suitability, species distribution, and the effects of climate change.

Data Availability Statement

The CTD bottom temperature and salinity data were downloaded from the NOAA/NEFSC Oceanography Branch Hydrographic Database: https://catalog.data.gov/dataset/oceanography-branch-hydrographic-database. The original and bias-corrected ROMS daily bottom temperature and salinity simulations are available at: https://doi.org/10.6084/m9.figshare.14245796.

Acknowledgments
The authors would like to thank Paula Fratantoni for pointing out the possibility of spatial biases in oceanographic models and four anonymous reviewers for improving the manuscript. This work was supported by an NOAA Fisheries and the Environment (FATE) grant, with funds provided through NOAA/CINAR (NA14OAR4320158).

References

Aung, M. T., Shrestha, S., Weesakul, S., & Shrestha, P. K. (2016). Multi-model climate change projections for Belu river basin, Myanmar under representative concentration pathways. *Journal of Earth Science and Climatic Change*, *7*(1), 1–13.

Buschow, S., Pidstrigach, J., & Friederichs, P. (2019). Assessment of wavelet-based spatial verification by means of a stochastic precipitation model (wv_verif v0.1.0). *Geoscientific Model Development*, *12*(8), 3401–3418. https://doi.org/10.5194/gmd-12-3401-2019

Cannon, A. J. (2016). Multivariate bias correction of climate model output: Matching marginal distributions and intervariable dependence structure. *Journal of Climate*, *29*(19), 7045–7064. https://doi.org/10.1175/jcli-d-15-0679.1

Chang, J. H., Chen, Y., Holland, D., & Grabowski, J. (2010). Estimating spatial distribution of American lobster *Homarus americanus* using habitat variables. *Marine Ecology Progress Series*, *420*, 145–156. https://doi.org/10.3354/meps08849

Chang, J. H., Shank, B. V., & Hart, D. R. (2017). A comparison of methods to estimate abundance and biomass from belt transect surveys. *Limnology and Oceanography: Methods*, *15*(5), 480–494. https://doi.org/10.1002/lom3.10174

Chen, J., Brissette, F. P., Chaumont, D., & Braun, M. (2013). Finding appropriate bias correction methods in downscaling precipitation for hydrologic impact studies over North America. *Water Resources Research*, *49*(7), 4187–4205. https://doi.org/10.1002/wrcr.20331

Chen, Z., Curchitser, E., Chant, R., & Kang, D. (2018). Seasonal variability of the cold pool over the Mid-Atlantic bight continental shelf. *Journal of Geophysical Research: Oceans*, *123*(11), 8203–8226. https://doi.org/10.1029/2018jc014148

Chen, Z., & Curchitser, E. N. (2020). Interannual variability of the Mid-Atlantic bight cold pool. *Journal of Geophysical Research: Oceans*, *125*(8), e2020JC016445. https://doi.org/10.1029/2020jc016445

Cressie, N. (1986). Kriging nonstationary data. *Journal of the American Statistical Association*, *81*(395), 625–634. https://doi.org/10.1080/01621459.1986.10478315

Giorgi, F. (2019). Thirty years of regional climate modeling: Where are we and where are we going next? *Journal of Geophysical Research: Atmospheres*, *124*(11), 5696–5723. https://doi.org/10.1029/2018JD030094

Gudmundsson, L. (2016). *qmap: Statistical transformations for post-processing climate model output*. [Computer software manual]. (R package version 1.0-4).

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., & Engen-Skaugen, T. (2012). Technical Note: Downscaling RCM precipitation to the station scale using statistical transformations – A comparison of methods. *Hydrology and Earth System Sciences*, *16*(9), 3383–3390. https://doi.org/10.5194/hess-16-3383-2012

Guisan, A., Edwards, T. C., Jr, & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecological Modelling*, *157*(2–3), 89–100. https://doi.org/10.1016/s0304-3800(02)00204-1

Gutiérrez, J. M., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., et al. (2019). An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the value perfect predictor cross-validation experiment. *International Journal of Climatology*, *39*(9), 3750–3785. https://doi.org/10.1002/joc.5462

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models. Vol. 43*. CRC Press.

Hengl, T. (2009). *A practical guide to geostatistical mapping*. Netherlands: Hengl Amsterdam.

Hiemstra, P. H., Pebesma, E. J., Twenhöfel, C. J. W., & Heuvelink, G. B. M. (2009). Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. *Computers & Geosciences*, *35*(8), 1711–1721. https://doi.org/10.1016/j.cageo.2008.10.011

Johnson, F., & Sharma, A. (2012). A nesting model for bias correction of variability at multiple time scales in general circulation model precipitation simulations. *Water Resources Research*, *48*, W01504. https://doi.org/10.1029/2011wr010464

Kang, D., & Curchitser, E. N. (2013). Gulf stream eddy characteristics in a high-resolution ocean model. *Journal of Geophysical Research: Oceans*, *118*(9), 4474–4487. https://doi.org/10.1002/jgrc.20318

Kang, D., & Curchitser, E. N. (2015). Energetics of eddy-mean flow interactions in the Gulf stream region. *Journal of Physical Oceanography*, *45*(4), 1103–1120. https://doi.org/10.1175/jpo-d-14-0200.1

Lafon, T., Dadson, S., Buys, G., & Prudhomme, C. (2013). Bias correction of daily precipitation simulated by a regional climate model: A comparison of methods. *International Journal of Climatology*, *33*(6), 1367–1381. https://doi.org/10.1002/joc.3518

Lazoglou, G., Angnostopoulou, C., Tolika, K., & Benedikt, G. (2020). Evaluation of a new statistical method-TIN-copula-for the bias correction of climate models' extreme parameters. *Atmosphere*, *11*(3), 243. https://doi.org/10.3390/atmos11030243

Li, B., Tanaka, K. R., Chen, Y., Brady, D. C., & Thomas, A. C. (2017). Assessing the quality of bottom water temperatures from the finite-volume community ocean model (FVCOM) in the northwest Atlantic shelf region. *Journal of Marine Systems*, *173*, 21–30. https://doi.org/10.1016/j.jmarsys.2017.04.001

Lowen, J. B., Hart, D. R., Stanley, R. R. E., Lehnert, S. J., Bradbury, I. R., & DiBacco, C. (2019). Assessing effects of genetic, environmental, and biotic gradients in species distribution modeling. *ICES Journal of Marine Science*, *76*(6), 1762–1775. https://doi.org/10.1093/icesjms/fsz049

Mao, G., Vogl, S., Laux, P., Wagner, S., & Kunstmann, H. (2015). Stochastic bias correction of dynamically downscaled precipitation fields for Germany through copula-based integration of gridded observation data. *Hydrology and Earth System Sciences*, *19*(4), 1787–1806. https://doi.org/10.5194/hess-19-1787-2015

Maraun, D. (2013). Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *Journal of Climate*, *26*(6), 2137–2143. https://doi.org/10.1175/jcli-d-12-00821.1

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

Maraun, D., Widmann, M., & Gutiérrez, J. M. (2019). Statistical downscaling skill under present climate conditions: A synthesis of the value perfect predictor experiment. *International Journal of Climatology*, *39*(9), 3692–3703. https://doi.org/10.1002/joc.5877

Maraun, D., Widmann, M., Gutiérrez, J. M., Kotlarski, S., Chandler, R. E., Hertig, E., et al. (2015). Value: A framework to validate downscaling approaches for climate change studies. *Earth's Future*, *3*(1), 1–14. https://doi.org/10.1002/2014ef000259

Matheron, G. (1962). *Traité de géostatistique appliquée 1* (Vol. 1). Editions Technip.

Mendez, M., Maathuis, B., Hein-Griggs, D., & Alvarado-Gamboa, L.-F. (2020). Performance evaluation of bias correction methods for climate change monthly precipitation projections over Costa Rica. *Water*, *12*(2), 482. https://doi.org/10.3390/w12020482

Odeh, I. O., McBratney, A. B., & Chittleborough, D. J. (1995). Further results on prediction of soil properties from terrain attributes: Heterotopic cokriging and regression-kriging. *Geoderma*, *67*(3–4), 215–226. https://doi.org/10.1016/0016-7061(95)00007-b

Pebesma, E., Cornford, D., Dubois, G., Heuvelink, G. B. M., Hristopulos, D., Pilz, J., et al. (2011). INTAMAP: The design and implementation of an interoperable automated interpolation web service. *Computers & Geosciences*, *37*(3), 343–352. https://doi.org/10.1016/j.cageo.2010.03.019

Pebesma, E. J. (2004). Multivariable geostatistics in S: The gstat package. *Computers & Geosciences*, *30*(7), 683–691. https://doi.org/10.1016/j.cageo.2004.03.012

R Core Team. (2020). *R: A language and environment for statistical computing [Computer software manual]*. Austria. Retrieved from https://www.R-project.org/

Ruybal, C. J., Hogue, T. S., & McCray, J. E. (2019). Evaluation of groundwater levels in the Arapahoe aquifer using spatiotemporal regression kriging. *Water Resources Research*, *55*(4), 2820–2837. https://doi.org/10.1029/2018wr023437

Shrestha, M., Acharya, S. C., & Shrestha, P. K. (2017). Bias correction of climate models for hydrological modeling – Are simple methods still useful? *Meteorological Applications*, *24*(3), 531–539. https://doi.org/10.1002/met.1655

Sunyer, M. A., Hundecha, Y., Lawrence, D., Madsen, H., Willems, P., Martinkova, M., et al. (2015). Inter-comparison of statistical downscaling methods for projection of extreme precipitation in Europe. *Hydrology and Earth System Sciences*, *19*(4), 1827–1847. https://doi.org/10.5194/hess-19-1827-2015

Tanaka, K. R., Chang, J.-H., Xue, Y., Li, Z., Jacobson, L., & Chen, Y. (2019). Mesoscale climatic impacts on the distribution of *Homarus americanus* in the U.S. inshore Gulf of Maine. *Canadian Journal of Fisheries and Aquatic Sciences*, *76*(4), 608–625. https://doi.org/10.1139/cjfas-2018-0075

Teutschbein, C., & Seibert, J. (2012). Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *Journal of Hydrology*, *456–457*, 12–29. https://doi.org/10.1016/j.jhydrol.2012.05.052

Teutschbein, C., & Seibert, J. (2013). Is bias correction of regional climate model (RCM) simulations possible for non-stationary conditions? *Hydrology and Earth System Sciences*, *17*(12), 5061–5077. https://doi.org/10.5194/hess-17-5061-2013

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. https://doi.org/10.1109/tip.2003.819861

Webster, R., & Oliver, M. A. (2007). *Geostatistics for environmental scientists*. John Wiley & Sons.
Wernli, H., Paulat, M., Hagen, M., & Frei, C. (2008). SAL – A novel quality measure for the verification of quantitative precipitation forecasts. *Monthly Weather Review*, *136*(11), 4470–4487. https://doi.org/10.1175/2008mwr2415.1
Wood, S. N. (2017). *Generalized additive models: An introduction with R*. CRC Press.
Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media.