Evaluating Current Statistical and Dynamical Forecasting Techniques for Seasonal Coastal Sea Level Prediction

Xiaoyu Long[a,b], Matthew Newman[b], Sang-Ik Shin[a,b], Magdelena Balmeseda[c], John Callahan[d], Gregory Dusek[d], Liwei Jia[e], Benjamin Kirtman[f], John Krasting[e], Cameron C. Lee[g], Tong Lee[h], William Sweet[i], Ou Wang[h], Yan Wang[a,b], Matthew J. Widlansky[j]

[a] *CIRES, University of Colorado Boulder, Boulder, Colorado, USA*

[b] *NOAA Physical Sciences Laboratory, Boulder, Colorado, USA*

[c] *European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom*

[d] *NOAA Center for Operational Oceanographic Products and Services, Silver Spring, Maryland, USA*

[e] *NOAA Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey, USA*

[f] *Rosenstiel School of Marine, Atmospheric, and Earth Science, University of Miami, Miami, Florida, USA*

[g] *Kent State University, Department of Geography, ClimRISE Laboratory, Kent, Ohio, USA*

[h] *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA*

[i] *NOAA National Ocean Service, Silver Spring, Maryland, USA*

[j] *Cooperative Institute for Marine and Atmospheric Research, School of Ocean and Earth Science and Technology, University of Hawai'i at Mānoa, Honolulu, Hawai'i , USA*

Corresponding author: Xiaoyu Long, xiaoyu.long@noaa.gov

ABSTRACT

The need for skillful seasonal prediction of coastal sea level anomalies (SLAs) has become increasingly evident as climate change has increased coastal flooding risks. Here, we evaluate nine current forecast systems by calculating deterministic and probabilistic skill from their retrospective forecasts ("hindcasts") over 1995-2015, for lead times up to 6-9 months, at two United States tide gauge stations (Charleston, SC and San Diego, CA). Additionally, we assess local skill enhancement by two post-processing/downscaling techniques: an observationally-based multivariate linear regression and a hybrid dynamical approach convolving sea-level sensitivity to surface forcings with predicted surface forcing variations. We find that all these approaches face challenges stemming from whether modeled SLAs sufficiently represent observed local coastal SLA variations, because of both ocean model limitations and inadequacies in model initialization and ensemble spread. Some of these issues also complicate the ability of the post-processing techniques to improve probabilistic skill, even though they do somewhat improve deterministic skill. In general, deterministic hindcast skill is considerably higher for San Diego than Charleston, as expected from the stronger influence of ENSO. However, ensemble spread metrics such as forecast reliability and sharpness remain low for both locations, highlighting model deficiencies in representing uncertainty. Additionally, evaluating how well any technique predicts seasonal coastal sea level variations is complicated by the forced trend component and particularly how it is estimated. Moreover, model skill is matched by a stochastically-forced multivariate linear prediction model constructed from observations, suggesting that substantial improvement remains for predicting coastal seasonal SLAs, which could also include leveraging other predicted fields including sea level pressure and prevailing winds.

SIGNIFICANCE STATEMENT

Coastal floodings have occurred more frequently in the last few decades, and it is anticipated that the number of such hazardous events will be increasing in the future. Therefore, accurate and reliable forecasting of coastal water level is becoming increasingly more important. This study thoroughly evaluated some current forecast techniques for sea level along the U.S. coasts and found that those techniques are still not capable to produce useable forecasting of anomalous sea level at U.S. east coast 3 months in advance, due to model inadequacy. The current generation of forecasting models were not designed for coastal sea level prediction, and we propose a few potential improvements that can potentially advance our capability in coastal sea level and inundation forecasting in the near future.

1. Introduction

Coastal flooding is a growing concern for the United States (U.S.) due to ongoing sea level rise (Church et al., 2013; Church & White, 2006; May et al., 2023), seasonal-to-decadal sea level variability, and land subsidence (Nicholls et al., 2021). These flooding events impact both ecosystems and public safety, with damages to the natural environment and built infrastructure including backed-up drainages, road closures, and saltwater intrusions. Flooding risks are projected to continue increasing in the coming decades (Taherkhani et al., 2020; Thompson et al., 2021; Vitousek et al., 2017), so there is an urgent need for accurate and reliable coastal flood prediction, ideally for months to seasons in advance. High coastal water levels are forced by a combination of drivers occurring at multiple spatial and temporal scales, including tides, waves, anomalous river runoff, storm surges, and sea level anomalies (SLAs) driven by atmospheric and oceanic processes on both weather and climate time scales. While the tide is usually the largest component of local water level variability, coastal flooding events typically occur when high tide coincides with other conditions favorable to anomalously high water levels. Though the astronomical (tide) contributions to sea levels are already well predicted, a high tide flooding prediction system should consider all these other factors as well (Hague et al., 2023).

Dusek et al. (2022) recently proposed a statistical approach to predict subseasonal to seasonal high tide flooding for U.S. tide gauges that are part of NOAA's National Water Level Observation Network (NWLON), which predicts daily probabilities of exceedance of a predefined flooding threshold for each location. These forecasts were made by combining tide predictions with a statistical representation of the "non-tidal residual" component of local sea levels, which includes both the long-term linear trend and the climatological distribution of hourly SLAs. They also showed that these forecasts were improved by including a simple damped-persistence model of monthly SLAs based on the observed autocorrelation function determined for each location. We might expect further improvement by using climate model seasonal SLA predictions that are more skillful than damped persistence.

Numerous studies have linked SLAs to patterns of seasonal climate variability (Han et al., 2017; Han et al., 2019b; Long et al., 2020; Roberts et al., 2016; Wang et al., 2023). This suggests that SLAs might be potentially predictable on seasonal time scales (Shin & Newman, 2021), and many studies have assessed seasonal SLA prediction skill of both dynamical and statistical models (Chowdhury et al., 2007; Long et al., 2023; McIntosh et al., 2015; Miles et al., 2014; Widlansky et al., 2017). Long et al. (2021) constructed a 10-model ensemble forecast and assessed its skill of forecasting SLAs with lead times up to 12 months, finding that the dynamical models produce more skillful forecasts than a damped-persistence model at most open ocean locations. However, these models generally do not have higher coastal skill than does an observationally-based Linear Inverse Model (LIM), a multivariate empirical dynamical model that also allows for transient anomaly growth (Shin and Newman (2021). Additionally, Frederikse et al. (2022) employed a hybrid dynamical approach, where observed surface forcings and predicted surface forcings, from hindcasts generated by state-of-the art seasonal forecast models, were projected onto SLA sensitivity at a specified location to global surface forcings computed by an ocean adjoint model. The resulting SLA hindcasts for the Charleston, SC location compared more favorably to observed tide gauge values there than the SLAs predicted by the same forecast model. Complicating all these skill assessments is that the pronounced externally-forced trend in sea level provides a substantial component of skill, at least as measured using commonly-used metrics (Wulff et al. 2022), that obscures the models' ability to predict seasonal climate variations (Long et al., 2021; Shin & Newman, 2021).

All the above studies of seasonal SLA prediction primarily focused on deterministic forecasts (e.g., ensemble means) and their skill assessment. However, warning end-users about high-tide flooding risks requires information about the likelihood of high-water (extreme) events (Dusek et al. 2022), which entails predicting tail probabilities. Therefore, it is also important to assess the probabilistic skill of coastal SLA prediction. For climate models, differences between multiple ensemble members (i.e., multiple forecast realizations) capture how initial uncertainty impacts the relative likelihood of future climate states. Probabilistic skill assessment then becomes a comparison, over the entire hindcast period, between the predicted probabilities of some extreme event and the actual chances of observing that event.

To improve our ability to forecast coastal flooding risk on seasonal and longer time scales, NOAA and NASA initiated the RISE project, a collaborative effort focused on developing and assessing novel dynamical and statistical forecast methods of SLAs along U.S. Coasts. This paper is an outgrowth of that project, which initially focused on a pilot study of monthly SLA forecast skill for sample tide gauge stations on the U.S. West and East Coasts (San Diego, CA and Charleston, SC). In this study, we evaluate monthly hindcasts of sea level anomalies for those two tide gauge stations using deterministic and probabilistic metrics. We also discuss challenges involved in making coastal SLA forecasts, including how trends in the model outputs impact skill assessment and how to use models that may not be correctly initialized with observed sea levels.

The paper is organized as follows. Section 2 reviews issues involved in making coastal sea-level predictions from the output of various dynamical and statistical models. Section 3 describes the forecast and observational datasets and the general skill metrics and methods used in this study. Section 4 presents the results of the deterministic and probabilistic skill assessment of seasonal SLAs, including a discussion of how this skill could be considerably impacted by both the presence of the externally forced trend and the inability of some hindcasts to represent it. Some remarks on how forecast models that are not initialized with satellite altimetry might be corrected follow in Section 5. Concluding remarks are made in Section 6, including recommendations for future advances in seasonal forecast systems to improve our prediction of coastal SLAs.

2. Challenges for Coastal Sea Level Seasonal Forecasts

As introduced above, previous studies have examined coastal SLA seasonal prediction skill. Yet it remains unclear how climate model output of monthly sea surface height anomalies should best be used to predict the risks of coastal flooding, and especially how these predictions should be verified against sea levels that are observed at tide gauges along the U.S. coastline. In most coupled climate models, the global ocean volume is conserved (i.e., they employ the Boussinesq approximation; Griffies & Greatbatch, 2012). As a result, these models cannot represent the global increase in sea level due to steric (thermal expansion) or barystatic (changes in water mass) processes, although they do allow for local height changes due to vertically integrated divergence/convergence, which is reflected in the model "sea surface height" (variable "zos" in the output from the Coupled Model Intercomparison Project (CMIP), as described in Griffies et al. (2016). This approach is sufficient for some purposes because changes in the global mean volume do not impact either ocean dynamics or coupling to the atmosphere. While the dynamical models considered here use a non-linear free surface, some older models use a "rigid lid approximation" with no variations in sea level at all; in this case, sea level must be calculated diagnostically from the model's ocean bottom pressure and density profiles (Griffies & Adcroft, 2008). In all cases, however, local density variations due to temperature and salinity changes can impact sea level locally, under the restriction that their global integral remains constant under the Boussinesq approximation.

The local change of sea level $\boldsymbol{\eta}$ can be expressed as (Griffies & Greatbatch, 2012):

$$\frac{\partial \eta}{\partial t} = \frac{Q_m}{\rho(\eta)} - \nabla \cdot \boldsymbol{U} - \int_{-H}^{\eta} \frac{1}{\rho} \frac{d\rho}{dt} dz, \tag{1}$$

where $\boldsymbol{Q_m}$ is water mass flux from the boundary (land ice melting, river runoff, etc.), $\boldsymbol{\rho(\eta)}$ is the water density at the surface, $\boldsymbol{H}$ is depth of the ocean, and $\boldsymbol{U}$ is the vertically integrated ocean current vector. The first term on the right-hand side of (1) is the contribution to the local sea level from mass input to the ocean, which is important to the inter-annual variability of global mean sea level (Hamlington et al., 2020). The second term is the vertically integrated divergence, accounting for the local change of dynamic sea level. The last term is the non-Boussinesq steric effect, which arises from density change following a fluid parcel and vanishes in a Boussinesq fluid. The global mean sea level change due to this non-Boussinesq steric effect can be corrected diagnostically as (Griffies & Greatbatch, 2012):

$$\eta(s,t) = \eta^B(s,t) + \frac{V_0}{A} \ln \frac{\rho(0)}{\rho(t)} \tag{2}$$

where $\boldsymbol{\eta(s,t)}$ is the sea level at a given location *s* at time *t*, $\boldsymbol{\eta^B(s,t)}$ is dynamic sea level from the ocean model, $\boldsymbol{V_0}$ is the initial reference volume of the global ocean, $\boldsymbol{A}$ is the global surface area of the ocean, $\boldsymbol{\rho(0)}$ is the initial global volume-averaged density, and $\boldsymbol{\rho(t)}$ is the global volume-averaged density at time *t*. To verify sea level forecasts against global altimetry observations and reanalyses, and ultimately predict coastal sea level, this quantity, $\boldsymbol{\eta(s,t)}$, needs to be predicted, but unfortunately many seasonal forecast systems typically output only $\boldsymbol{\eta^B(s,t)}$.

Global reanalyses that assimilate both in situ and satellite observations are used to initialize and verify seasonal forecasts produced by climate models. For sea surface height, global observations are only available since 1993, from satellite altimetry. Some ocean reanalysis systems also use altimetry to correct the global mean sea level change (Balmaseda et al 2013). However, not all seasonal forecast climate models include altimetry in their initialization. Long et al. (2021) noted that models that included assimilation of altimetry data in their initialization tend to have a more realistic trend (due to both internal variability and external forcing) in their hindcasts than models that did not, which complicated skill comparison between models, especially in regions of strong sea surface height trends such as the U.S. East Coast. Initialization that uses altimetry appears to improve seasonal SLA prediction in many ocean regions, although less obviously so along the North American coastline (Widlansky et al., 2023) where the benefit of altimetry observations may not have been fully realized in the present-generation of assimilation systems (Feng et al. 2024).

Complicating matters further is that station-based tide gauges measure the water level relative to benchmarks on land (Gill & Schultz, 2001; Pugh & Woodworth, 2014). Also, some gauges are in bays or inlets, which can complicate their relationship to coastal sea level, and their measurements may include effects of freshwater flows from upstream (Piecuch et al., 2018). Also, since the land itself may move over time, tide gauge measurements can implicitly include a component due to vertical land motion (VLM) (Wöppelmann & Marcos, 2016). While for seasonal forecasts VLM is so small that it can be neglected (and none of the regional or global models simulate VLM), it can become important over the entire multi-decade period common to most seasonal hindcast datasets, where it is not easily accounted for (Ray et al., 2023; Zervas et al., 2013).

Seasonal climate model forecasts are often "mean bias-corrected", a post-processing step in which potentially erroneous model climatological mean states are replaced with the observed climatological mean state, which typically depends upon both the seasonal cycle and the forecast lead time (Stockdale et al 1993). That is, forecasts are verified by comparison of observed anomalies (relative to the observed mean state) to predicted anomalies (relative to the model mean state at that lead time). Typically, mean states are defined over a few decades, long enough to reduce sampling effects but still reasonably short enough to be representative of the current climate state in the context of long-term (i.e., centennial scale) climate change. Unfortunately, as climate warming has accelerated over the latter half of the 20th century, this latter assumption is not valid for SLAs at many locations, especially along the East and Gulf coasts that have experienced an accelerating trend in mean sea levels over the past few decades relative to the global mean (Hamlington et al., 2020). The presence of a trend, even over a relatively short climatological period, leads to two issues. First, as noted above, the forecast model may not be able to entirely simulate all the processes responsible for the sea level trend itself. Second, anomalies are typically defined relative to a fixed long-term mean, which means that the trend component is included as part of the anomaly and, therefore, has a pronounced impact on the estimation of seasonal skill (e.g., Wulff et al., 2022). This is illustrated in Fig. 1 by considering a simple case where hindcasts of seasonal variations are so unskillful that they are entirely uncorrelated with the observed time series, but where both hindcasts and observations are also superposed about a common linear trend. Then, the resulting two time series would be well correlated (Fig. 1a), making the seasonal forecast system appear skillful. Conversely, an incorrect trend in the hindcasts relative to observations could reduce skill even where the seasonal variations of the hindcasts and observations were otherwise well correlated (Fig. 1b). In cases such as illustrated in Fig. 1, we might simply remove or correct the linear trend. For example, Balmaseda et al (2024) show that a simple linear trend correction adds skill to seasonal SLA forecasts. More generally, however, evaluating the impact the trend on local hindcast skill becomes problematic when the externally-forced trend is nonlinear, particularly for relatively short records when it is unclear how to separate the trend from natural variability (e.g., Solomon et al. 2011).

3. Data and Methods

Here, we discuss how we assess hindcast skill from various seasonal forecast systems, encompassing purely dynamical models, purely statistical models, and hybrid techniques. Since previous studies of (deterministic) sea level forecast skill all used different hindcast periods, we assess skill of all these techniques for 1995-2015, which is the common period for hindcast availability from all the forecast techniques.

3.1 Tide gauge verification data

The verification data are based on monthly mean sea level data at the San Diego and Charleston NOAA NWLON tide gauges from 1995 to 2016, obtained from the Permanent Surface for Mean Sea Level (PSMSL; Holgate et al., 2013). SLAs are defined by removing the 21-year monthly mean climatology from each tide gauge time series. We limited our skill assessment to these two stations since the adjoint model of the Estimating Circulation and Climate of the Ocean (ECCO) system had only developed hindcasts there.

While we focus on SLA prediction at the tide gauges, we also discuss results from three other observationally-based gridded datasets: the SSALTO/DUACS multimission satellite altimetry dataset (Hauser et al., 2021; also known as AVISO in the literature) with a 1/4º spatial resolution, and three ocean reanalyses, ORAS5 (Zuo et al., 2019) with a 1/4º spatial resolution, ECCO (Forget et al., 2015) with 1/4º spatial resolution, and GLORYS12 (Jean-Michel et al., 2021) with a 1/12º spatial resolution. As discussed in Section 2, the tide gauges measure quantities beyond what may be captured by global oceanic datasets, which is illustrated by comparing the gauge time series with the nearest grid value from the gridded datasets, shown in Fig. 2. Also shown is the correlation of the monthly gauge-located SLAs of each of the gridded datasets and the tide gauge time series. All the gridded products capture the San Diego tide gauge reasonably well, but ORAS5 and ECCO capture only about half the monthly SLA variance observed at the Charleston tide gauge, which may be partly due to the relatively low weight that coastal data is given in their data assimilation systems (Feng et al. 2024). Finally, we also show linear and quadratic trend lines to each tide gauge record, determined by fitting a line or quadratic curve, respectively, to the data by minimizing the mean squared error. Note that there appears to be some upward trend over the period of record, which seems to have accelerated for both gauges after about 2011 when global mean sea surface temperatures also began to increase more rapidly (Garcia-Soto et al., 2021), so it is unclear how well either least-squares fit captures the externally-forced trend component.

3.2 Hindcast techniques and data

First, we consider hindcasts from three traditional assimilation-initialized seasonal forecast systems based on coupled dynamical models: CCSM4 (Community Climate System Model Version 4, Kirtman et al., 2014), SPEAR (Seamless System for Prediction and Earth System Research, Delworth et al., 2020; Lu et al., 2020), and ECMWF SEAS5 (Johnson et al., 2019). Apart from other modeling framework differences, including horizontal resolution (see Table 1), these forecast systems differ in how the ocean state is initialized and how the ocean model simulates global mean sea level evolution (although note that all the ocean models have a free surface):

(1) CCSM4 has the Parallel Ocean Program version 2 (POP2) model as its ocean component and is initialized with the Climate Forecast System Reanalysis (CFSR). Though CFSR captures the realistic variation of ocean heat content and hence the variation of SLA (Xue et al., 2011), the POP2 model by construction requires the global mean sea level to remain constant. Consequently, CCSM4 has no trend in its global mean sea level. Also, CCSM4 does not account for global mean variations in freshwater fluxes.

(2) SPEAR uses its ocean data assimilation to initialize its ocean model, the Modular Ocean Model Version 6 (MOM6). This data assimilation incorporates observed temperature and salinity profiles from ARGO based on the Ocean Tendency Adjustment (OTA) outlined in Liu et al. (2020). SPEAR does not explicitly simulate the global mean steric sea level evolution from internal changes in heat and salt due to the previously discussed limitations of a Boussinesq model. Still, unlike CCSM4, it does consider the imbalance of the hydrological

cycle of the climate system, which is accounted for within the topmost layer of the ocean model (Cazenave et al., 2012).

(3) SEAS5 also uses a Boussinesq ocean model, the Nucleus for European Modelling of the Ocean (NEMO) model. Unlike SPEAR, however, the ocean initial conditions of SEAS5 include information from the altimeter observations via assimilation, including the global steric change (Zuo et al., 2019). Therefore, the SEAS5 forecasts of sea level will inherit the information from the sea level trend in their initial conditions and have a more realistic sea level trend both regionally and globally.

The impact of some of these configuration differences in each forecast system is seen when comparing observationally-based monthly anomalies of global-mean sea level, determined from the ORAS5, GLORYS12, and AVISO datasets (Section 3.1), to globally averaged lead-1 month SLAs from the CCSM4, SPEAR, and SEAS5 hindcasts (Fig. 3). [Note that here we define the lead-1 month as the first month after initialization, so that it is a combined representation of the initial climate state and the short-term model evolution from it.] Figure 3a shows that while there are some differences between the observationally-based time series, all three capture the long-term trend in the global mean, as do the SEAS5 lead-1 hindcasts. The observed evolution is not captured by the lead-1 hindcast anomalies output from either the SPEAR or the CCSM4 (Fig. 3b), which, as noted above, do not include the global-mean steric or barystatic components. For the SPEAR, following the approach discussed in Section 2, the global-mean steric component was computed from one ensemble member of the lead-1 hindcasts and added to the global mean lead-1 ensemble-mean hindcast (Fig. 3b) to produce the "SPEAR+steric" curve in Fig. 3a, yielding a closer match to observations. Still, there remains a discrepancy, likely due to the lack of information about changes in initial oceanic volume that assimilation of satellite altimetry could provide. Finally, comparing the linearly detrended global-mean sea level anomalies for the three lead-1 hindcast datasets to that determined from ORAS5 (Fig. 3c) shows the CCSM4 and SPEAR initializations may not entirely capture interannual variations in the global mean sea level, even apart from the trend.

We also created a downscaled version from each of the dynamical forecast ensembles, using the technique demonstrated in Long et al. (2023): A seasonally invariant deterministic downscaling operator was constructed by multivariate linear regression of the high-resolution (1/12°) GLORYS12 ocean reanalysis data against its coarse-grained (1°) counterpart. Then, the downscaling operator is applied to each ensemble member of each dynamical hindcast, generating an ensemble of high-resolution coastal sea level hindcasts that we refer to as DownscalingCCSM4, DownscalingSPEAR, and DownscalingSEAS5, respectively. By downscaling each ensemble member rather than the overall ensemble-mean (as was done in Long et al. (2023)), we generate a downscaled hindcast ensemble whose spread is based upon the original model ensemble. Note that the downscaling operator is in a reduced Empirical Orthogonal Function (EOF) space, so that not all the variance of the original hindcasts is retained in the downscaled hindcasts. Multi-model ensemble means were constructed using either the hindcast ensembles from the three GCMs or their corresponding downscaled hindcast ensembles.

Frederikse et al. (2022) developed a hybrid dynamical approach for seasonal SLA prediction. They first computed the sensitivities of the coastal sea level at a specific location to different global atmospheric surface forcings, using the ECCO adjoint model. Then, SLA prediction is made by convolving these sensitivities to observed and predicted atmospheric surface forcings, made up of observed (ECCO) forcings up to 12 months prior to initialization time followed by predicted atmospheric forcings up to 12 months after initialization time. Note that the precise length of applied forcings depends upon forecast lead time (e.g., for a 5-month lead, the forcing consists of 12 months of observed forcing followed by 5 months of predicted forcing). In Frederikse et al. (2022), the predicted atmospheric forcing fields are from a 10-member CCSM4 model. In the present study, we also use a 15-member SPEAR model in addition to the 10-member CCSM4 model for the predicted surface forcings. The resulting hindcast ensembles are named ECCO_CCSM4 and ECCO_SPEAR, respectively. Hereafter, this approach is referred to as the ECCO adjoint approach.

Note that for both the CCSM4 and SPEAR, we assess the skill of the original dynamical hindcasts, the downscaled version of those hindcasts, and the ECCO adjoint model forced by those hindcasts (albeit using predicted atmospheric surface forcing variables rather than predicted sea surface heights). This yields an ideal suite of forecasts to compare each method because they are all derived from the same dynamical forecast system (CCSM4 or SPEAR).

Finally, we also included hindcasts from a LIM, trained using near-global gridded fields of SST from HadISST (Kennedy et al., 2019) and SLA from ORAS4 (Balmaseda et al., 2013) from 1961 to 2015 (Shin & Newman, 2021). The LIM's deterministic forecast is represented by its ensemble mean, and its fixed but lead-dependent expected error statistics are used to estimate the uncertainty (i.e., ensemble spread) of its forecasts (equation (8) in Penland & Sardeshmukh, 1995). The LIM hindcasts were ten-fold cross-validated for the entire 1961-2015 period, but for this paper, we assess skill only for those hindcasts initialized in the common 1995-2015 period.

Similar to earlier studies (Frederikse et al., 2022; Long et al., 2021), we use a univariate AR1 model, or damped persistence (van den Dool, 2006), as a minimum baseline of skill for all the evaluated forecast techniques. Note that LIM and damped persistence are similar in that both are determined from the lead-1 autocovariance of the data, but the LIM yields a multivariate matrix operator rather than a univariate scalar, so it also yields transient anomaly growth that leads to additional state-dependent predictability (Shin & Newman, 2021). Like the LIM, an AR1 model has fixed but lead-dependent expected error statistics, which can be used to estimate its prediction uncertainty. The damped-persistence coefficients were determined from each tide gauge record, using data only during the hindcast period, and cross-validated using a leave-one-out methodology.

Most dynamical models are initialized with a near-instantaneous or daily field; this is on the first day of the month for the three dynamical forecast systems assessed here. The first monthly mean forecast is then the mean of the first month of the forecast run, sometimes called the “Month 0.5” forecast, i.e., centered in the middle of the calendar month (e.g., Kirtman et al., 2014). In contrast, empirical models may be initialized with observed (tide gauge/reanalysis) monthly mean anomalies centered on the previous month, so that the 1-month lead LIM/damped-persistence forecast and the dynamical model Month 0.5 forecast verify simultaneously. Following Newman and Sardeshmukh (2017), for clarity we renamed both these forecasts the “Month 1” forecast (i.e., the first month of the forecast period), and so on for increasing forecast leads (see also schematic in Ding et al., 2018).

For the three dynamical forecast systems (CCSM4, SPEAR, and SEAS5) that are initialized using full-field variables, a mean bias correction is first applied by removing the lead-time dependent climatology determined during 1995--2015 (Smith et al., 2013), as discussed in Section 2. The statistical downscaling is applied to these bias-corrected anomaly fields. LIM hindcast anomalies, initially defined relative to the 1960-2015 period, are adjusted to be relative to the 1995-2015 climatology, but otherwise are uncorrected. The ECCO_CCSM4 and ECCO_SPEAR each are forced with the mean bias-corrected dynamical model atmospheric forcing ensemble members from CCSM4 and SPEAR hindcasts, respectively, with an additional mean bias-correction applied to the resulting adjoint model hindcasts.

3.3 Prediction skill metrics

Deterministic skill is assessed using the anomaly correlation coefficient (ACC) between observations and ensemble-mean predictions, as a function of lead time, either computed over all calendar months or calculated separately for each verification calendar month. ACC measures how well a model can predict the phase and sign of observed anomalies (Wilks, 2011). We also computed the root-mean-squared (RMS) skill score (RMSSS; e.g., Newman and Sardeshmukh (2017)), defined as $\varepsilon \equiv 1 - \hat{\sigma}$, where the standardized error $\hat{\sigma} = \sigma/\sigma_{obs}$, σ is the RMS forecast error between observations and ensemble-mean predictions, and σ_{obs} is the observed climatological RMS value. RMSSS is a measure of the average relative amplitude of the forecast error, defined so that a perfect forecast has RMSSS=1, a climatological forecast (i.e., a predicted anomaly of zero) has RMSSS=0, and a forecast poorer than climatology has a negative score.

Probabilistic skill is assessed using two different metrics. First, we determined reliability diagrams (Weisheimer & Palmer, 2014), where hindcasts are grouped into bins according to the predicted probability (horizontal axis), and then plotted against the frequency at which observed events occur (vertical axis). For a perfectly reliable forecast system, predicted probabilities should match observed probabilities, in which case the reliability curve lies along the diagonal: If an event is predicted as having an x% probability of occurring, then the event should occur x% of the time. Also included are “sharpness” diagrams, showing how often each forecast probability is issued, particularly distinguishing forecasts other than the climatological probability (Wilks, 2011). For example, for a three-category tercile forecast, a sharp forecast system should be able to issue forecast probabilities other than the climatological probability of 0.33. We also calculate the reliability value as a single metric for easy comparison across techniques (Toth et al., 2006).

Finally, ROC (Receiver Operating Characteristic; Kharin & Zwiers, 2003) curves were constructed by plotting the false alarm rate against the hit rate for different probability thresholds. In general, as we lower the probability thresholds, more ‘positive’ forecasts will be issued, and hence, both the hit rate and the false alarm rate will increase. A good forecast system has a high hit rate while minimizing false alarms, so an ideal ROC curve is away from the diagonal towards the upper left corner of the diagram (Mason & Graham, 1999), thereby maximizing the area under the ROC curve (ROC area). ROC skill score (ROCS), defined as ROCS = 2(ROC area - 0.5), measures this quantity. ROCS values can range from -1 to +1, where ROCS<0 indicates skill worse than climatology (i.e., random chance).

4 Hindcast skill

Before we show the skill evaluation, in Figure 4 we display an example of hindcast ensembles from each of the techniques for both San Diego and Charleston, initialized as of August 1, 1997 (the LIM is initialized using the monthly anomalies of July 1997) with lead times up to 12 months. In this case, August 1997 is Month 1, September 1997 is Month 2, and so on.

The anomalous sea level at San Diego was heavily influenced by that year's strong El Niño event and the associated coastally trapped Kelvin wave that propagated along the west coast of North America from the Tropics (Hamlington et al., 2015; Ryan & Noble, 2002). This led to an observed SLA maximum of nearly 18 cm in San Diego by November. While the models had some hint of this coastal Kelvin wave, it was too weak and delayed by a couple of months (Balmaseda et al., 2002), which resulted in a relatively flat SLA response, even for the individual ensemble members. This error was also apparent for hindcasts initialized earlier in June and July, and it was not until the October initialization that the models captured the timing of the November maximum (not shown, but see https://www.psl.noaa.gov/forecasts/SeaLevel/#RISE for all hindcasts and verifications used in this paper). Interestingly, the three models also appeared to predict a similar delayed and too-weak SLA response in San Diego for the 2009-10 and 2015-16 El Niño events (not shown). The ensemble spread of the downscaled and ECCO adjoint approaches was mostly reduced compared to the original models' spread. The observed November maximum was not included within any technique's ensemble spread, including the LIM's 2 standard-deviation ensemble spread (shading).

For Charleston, all nine techniques predicted a flat SLA response with increasing lead. Notably, the observed February 1998 SLA maximum of 20 cm was not contained within the ensemble spread of any forecast technique. There is a striking difference between the observed tide gauge value and the Month 1 forecast for SPEAR, DownscalingSPEAR, SEAS5, and DownscalingSEAS5. The ECCO adjoint approach appears to have improved the comparison between Month 1 and the corresponding observed tide gauge monthly mean anomaly. Finally, note ECCO_SPEAR appears to reduce ensemble spread relative to SPEAR, but the opposite is true for ECCO_CCSM4 compared to CCSM4.

4.1 Deterministic skill

The deterministic skill of all the techniques for the common hindcast period (Fig. 5) is considerably higher at San Diego than at Charleston, in agreement with previous studies (Long et al., 2021; Shin & Newman, 2021). For San Diego, SEAS5 and DownscalingSEAS5 have the highest skill, exceeding the multi-model ensemble mean (MMM) skill largely because the CCSM4 skill is poor. The LIM has skill comparable to MMM, SEAS5, and SPEAR for shorter leads, but its skill degrades faster for longer leads. The downscaling technique improves of CCSM4 but not SPEAR skill, consistent with Long et al. (2023). The ECCO adjoint approach improves upon CCSM4 skill even more than the downscaling technique but worsens SPEAR skill. The multi-model mean of the downscaled hindcasts has a similar skill to that of the dynamical model hindcasts. Note that for Months 1-2, only SEAS5 and the LIM have a skill that exceeds damped persistence (gray background shading), but as lead time increases, more techniques show relatively greater skill.

At Charleston, in contrast, the LIM has the highest skill, with SEAS5 having the highest skill of the dynamical models, slightly exceeded by the multi-model mean at longer leads. The ECCO adjoint approach improved the skill of both models, especially the CCSM4, while the downscaling technique slightly improved the skill of SPEAR but not of CCSM4.

As discussed in Section 2, hindcast skill evaluation is complicated by the existence of a pronounced externally-forced sea level trend, which has pronounced regional variations (e.g., cf. San Diego and Charleston in Fig. 2) that also are captured differently by the initializations of the different forecast techniques. For example, the higher Charleston skill for SEAS5 and the LIM might be due to their more accurate initialization of the externally-forced trend, although recall from Fig. 2 that the ORAS5 (and likewise the ORAS4, which the LIM is trained upon) still has some important differences from the tide gauge in Charleston. To evaluate the impact of this trend on hindcast skill, however, we would need to first distinguish it from natural internal climate variability, which is complicated by the short observational dataset (e.g., Deser et al., 2014; Frankignoul et al., 2017). In turn, whether the externally-forced trend is linear or nonlinear, which is unclear from Fig. 2, can impact estimates of internal variability.

How to precisely determine the externally forced trend is beyond the scope of this study (although see discussion in Shin & Newman, 2021), so here we instead test the sensitivity of our skill assessment upon the two different trend estimates shown in Fig. 2. Specifically, we first remove either the linear trend or the quadratic trend from both hindcasts and verification data, determined separately for each, and then recompute skill metrics of the resulting detrended data. The results for linear (quadratic) detrending are shown in Figs. 5cd (Figs. 5ef). Additionally, since a forced trend could contribute to persistence, we recomputed the damped-persistence models separately for each detrending.

All the techniques have considerably reduced skill for the detrended hindcasts (verified against detrended observations) at Charleston, consistent with earlier studies (Long et al., 2021; Shin & Newman, 2021), suggesting that much of the apparent skill on the U.S. East Coast is due to the pronounced trend there over the past three decades (Han et al., 2019a). That is, much of the Charleston skill is not from hindcasts that capture month-to-month variations of sea level but rather arises because both hindcasts and observational anomalies have a large trend component (relative to the constant climatological 1995-2015 mean) that artificially inflates estimates of the prediction skill of monthly variations (e.g., Fig. 1a). Note that while the qualitative impact of detrending on skill is similar for all the techniques, its greatest impact occurs for those techniques with relatively realistic initializations including realistic trends. However, determining the quantitative impact of the trend on skill seems sensitive to the assumed form of the trend. For example, linear detrending (Fig. 5d) causes a larger decrease in SEAS5 skill than does quadratic detrending (Fig. 5f). On the other hand, the skill of both SPEAR and the ECCO adjoint approaches decreases more for quadratic detrending. Still, for both detrending methods, the skill of many of the forecast techniques exceeds damped persistence even at longer forecast leads. Finally, note that while linear detrending has a much less pronounced impact on hindcast skill for San Diego, its impact is not negligible, especially for some techniques and with increasing leads.

Using RMSSS as the deterministic skill metric rather than ACC (Fig. 6) gives a similarly qualitative picture for San Diego skill, including the relative ordering of skill across the techniques and the impact of detrending. However, for Charleston, with the RMSSS metric, there are now fewer techniques whose skill exceeds damped persistence, although the relative position of the LIM is unchanged; after detrending, only a few techniques have skill that even matches damped persistence, with RMSSS that is only slightly positive. Additionally, it is apparent in Fig. 6 that when the dynamical models' ACC goes below about 0.4, RMSSS becomes negative, indicative of skill worse than a fixed prediction of a zero anomaly (i.e., climatology). However, this is not the case for either the LIM or damped persistence; for example, for damped persistence, RMSSS approaches zero only as ACC approaches zero. Interestingly, for single-member forecast systems, ACC=0.4 is equivalent to 100% standardized error (Livezey & Chen, 1983).

Hindcast skill for both stations depends upon the target month, especially for San Diego (Fig. 7). This result was found in some previous studies (Long et al., 2023; Shin & Newman, 2021) as well, but here the seasonality of skill for all the forecast techniques is compared together for a common period. For San Diego, skill maximizes for forecasts verifying during winter, which could be associated with the El Niño-Southern Oscillation (ENSO) signal that typically also has a wintertime maximum along the West Coast (e.g., Shin & Newman, 2021). Some techniques also show skill for up to 1-2 season leads when verifying during summer, consistent with an ENSO signal that is not predictable until after spring (the "spring predictability barrier"; e.g., Tippett & L'Heureux, 2020). Interestingly, the ECCO adjoint approach especially improves skill during summer, particularly for the CCSM4, whose skill is otherwise considerably worse than the other models. In contrast, changes in skill from downscaling had a much weaker seasonal dependence. Still, neither ECCO_CCSM4 nor ECCO_SPEAR summertime skill exceeds that of the LIM or SEAS5. SEAS5 shows significant skill up to Month 7 throughout the year. Linear and quadratic detrending have quantitatively similar effects to those in Fig. 5, mainly for spring and summer verifications when skill is already lower (not shown).

For Charleston, where skill is generally much lower than for San Diego, there is a less clear impact of seasonality upon skill (Fig. 8). Some of the techniques (DownscalingCCSM4, SPEAR, and DownscalingSPEAR) only have significant positive ACC during the early winter months. In contrast, SEAS5 and DownscalingSEAS5 have significant skill during spring and somewhat during fall, even at the longest leads available, with the LIM having a similar pattern with generally higher ACC values. Again, the ECCO adjoint approach boosts skill during the warm season, yielding significant skill for April to August verifications through Month 5 that exceeds (albeit not significantly) the skill from any other techniques. Interestingly, much of the ECCO adjoint skill increase for Charleston is retained even after detrending (Fig. 9). However, the degree of improvement is less when the trend line is quadratic (not shown) rather than linear.

4.2 Probabilistic skill

4.2.1 Reliability and sharpness

As was the case for deterministic skill, probabilistic metrics are better for San Diego (Fig. 10) than for Charleston (Fig. 11). Reliability is generally better for predictions of upper than lower tercile events. The most reliable forecasts are made by the LIM and SEAS5, even for the lower tercile events. For San Diego, the LIM is more reliable than the other techniques at this lead (Month 4), even as its deterministic skill is relatively poorer. For Charleston, except for the LIM, none of the hindcasts are particularly reliable, and even the LIM is more reliable for the upper than the lower tercile. The ECCO adjoint technique also has minimal impact on reliability, slightly improving CCSM4 but making SPEAR worse, notably for the lower tercile. The multi-model means do not improve overall reliability at either location (not shown) and slightly degrade it for the downscaled hindcasts. However, this may be due to the small number of models we used.

The inset sharpness diagrams show that while all the models can issue forecast probabilities other than the climatological value of 0.33, these hindcasts are dominated by very low forecast probabilities (i.e., they are generally clustered in the leftmost bins). Interestingly, SPEAR and SEAS5 tend to have more forecasts with a higher probability for Charleston than for San Diego (note their U-shape sharpness diagrams). Finally, note that the adjoint technique reduces the occurrence of higher forecast probabilities (e.g., 0.7 and 0.9 bin) compared to the models upon which they are based, particularly for Charleston, consistent with the reduced ensemble spreads for the ECCO_CCSM4 and ECCO_SPEAR seen in Fig. 4.

After linearly detrending Charleston hindcasts and observations, hindcasts become much less reliable (Fig. 12). That is, much of the (limited) reliability seen in Fig. 11 represents probabilities from hindcasts either early in the hindcast period when both hindcast and observed anomalies include a trend component that is relatively large and negative (when defined as an anomaly relative to the long-term mean), or late in the period when that trend component is relatively large and positive. It is not entirely surprising that the reliability and sharpness of the techniques with more realistic initialization (SEAS5 and LIM) are most impacted by the detrending. As with the deterministic metrics, the impact of the trend on reliability is much less at San Diego. Additionally, reliability is less sensitive than ACC to removing a quadratic rather than a linear trend, at least for the three-category approach used here (not shown).

4.2.2 ROC skill scores

The ROC curves for San Diego for Month 4 (Fig. 13) show that all the techniques have better performance in predicting upper tercile than lower tercile events. SEAS5, Downscaling SEAS5, and LIM have the best performance, with ROCS values between 0.78 to 0.82 for the upper tercile and 0.65 to 0.77 for the lower tercile, followed in order by the SPEAR and then the CCSM4. The downscaling and ECCO hybrid methods improve probabilistic skill for CCSM4 but not so much for SPEAR. The multi-model mean improves upon CCSM4 and SPEAR (and their related hindcasts) for both upper and lower tercile hindcasts but does not improve upon SEAS5. Note also that many techniques have relatively low hit rates and false alarm rates even for the lowest classification threshold because the techniques do not issue enough ‘positive’ event predictions overall.

For Charleston at Month 4 (Fig. 14), ROCS values for upper and lower terciles are generally lower than for San Diego. The LIM, SEAS5, and DownscalingSEAS5, in that order, have the highest ROCS values. The CCSM4 and its derived models (DownscalingCCSM4 and ECCO_CCSM4) all have ROCS near zero, indicating that the CCSM4 has no skill compared to a random classification model. The SPEAR and its derived models are only slightly better. The multi-model mean also again improves upon CCSM4 and SPEAR but not SEAS5.

These results are representative of other forecast lead times, which is demonstrated by the ROCS values for each of the techniques as a function of lead (Figs. 15 and 16). For San Diego (Fig. 15), SEAS5, DownscalingSEAS5, and LIM have the highest ROCS values through Month 7. The CCSM4 and its derived hindcasts have the lowest ROCS values, with SPEAR in between. Neither downscaling nor ECCO hybrid methods improve ROCS values of SEAS5 and SPEAR, although the ECCO hybrid does improve CCSM4. For Charleston (Fig. 16), the overall skill scores are lower than for San Diego, where again, the highest ROCS values are for SEAS5, DownscalingSEAS5, and LIM. Similarly, the downscaling technique does not improve the skill score, while the ECCO hybrid method slightly improves the skill over shorter lead times.

The removal of the linear trend also effectively reduces ROCS for SEAS5 and DownscalingSEAS5, but less so for the LIM (Fig. 16). Detrending improves the ROCS of CCSM4 and DownscalingCCSM4, mostly because the removal of the spurious trend in those model hindcasts improves the quality of their hindcasts, in the same way that the detrending degrades the hindcasts which have more realistic trends (Fig. 1b). In contrast to the deterministic skill, the impact of the trend upon probabilistic skill does not much depend upon whether the trend is estimated as quadratic or as linear (not shown).

5 Correcting forecasts from models with inadequate sea level initialization

As discussed in Section 2, using and evaluating dynamical model predictions of coastal SLAs can be challenging when the model output sea level variable does not entirely correspond to tide gauge measurements, particularly when the model is not initialized with sea level observations. Perhaps this could be alleviated by evaluating prediction skill after removing the global mean from both verifications and model hindcasts, addressing the absence of the steric contribution to global mean sea level in forecast models and/or their initializations. We considered this approach as a possible solution both to this problem and to the problem of evaluating detrended skill since a significant portion of the trend is related to global mean sea level rise (e.g., Fig. 3a), although for some, if not most, tide gauges the trend also has VLM contributions. Unfortunately, global mean sea level is also impacted by large scale internal variability (e.g., Fig. 3c), including potentially predictable climate variations such as ENSO (Cazenave et al., 2012; Wang et al., 2021). Hence, removing the global mean component was inadequate to comprehensively evaluate the skill of seasonal prediction of regional sea level anomalies. Likewise, removing basin-wide means was also unsuccessful.

Alternatively, we might assume that models whose sea level is incompletely initialized (e.g., without altimetry in the data assimilation) might be capable of predicting month-to-month sea-level changes so long as other ocean variable initializations (e.g., temperature and salinity) are not substantially impacted. Such an initialization error might be considered simply an offset, addressed by adjusting the hindcasts. Let the predicted monthly sea level state for initial time t and lead time j be $\tilde{\mathrm{Z}}(t, j)$. Then the prediction increment, or "delta", of sea level can be calculated from the model output as: $\Delta\tilde{\mathrm{Z}}(t, j) = \tilde{\mathrm{Z}}(t, j) - \tilde{\mathrm{Z}}(t, j-1)$, for all lead times. Finally, the adjusted forecast is determined by incrementing the observed initial monthly anomaly, $\mathrm{Z}(t, 0)$, with each delta at different lead times (i.e., $\mathrm{Z}(t, 0) + \Delta\tilde{\mathrm{Z}}(t, j)$), yielding a "delta-corrected sea level prediction" whose month-to-month change is identical to the original model forecast but is "initialized" with observations. However, this leaves us still with the choice of the initial observed monthly SLA. The most recent observed monthly sea level at the tide gauge is the previous month, or $\mathrm{Z}(t, -1)$. Using that as the initialization means that we still need $\Delta\tilde{\mathrm{Z}}(t, 0)$, which we determine from the difference between the current Month 1 and previous Month 1 forecasts ($\Delta\tilde{\mathrm{Z}}(t, 0) = \tilde{\mathrm{Z}}(t, 0) - \tilde{\mathrm{Z}}(t-1,0)$); then, $\tilde{\mathrm{Z}}(t, 0) = \mathrm{Z}(t, -1) + \Delta\tilde{\mathrm{Z}}(t, 0)$, and the forecasts can be incremented from that point onwards.

There are two benefits of this correction. First, the inconsistent trend between the model and observations is no longer an issue because the realistic trend is built into the corrected sea level. Second, the imperfect initialization is less of a problem since starting from the previous month's observed SLA will largely eliminate the difference between model initialization and observation.

The delta-corrected sea level prediction is particularly appealing for issuing real-time SLA predictions that are appropriate for specific tide gauges, especially when using models initialized without a trend component to predict SLAs for tide gauges with pronounced observed trends (which over a few decades could also include vertical land motion). Note that all delta-corrected tide gauge hindcasts include the observed trend after correction, allowing comparison with other techniques already initialized with observations. The delta-correction improves the ACC for the models without a correct trend, both the CCSM4 and the SPEAR, especially for San Diego (Fig. 17; cf. lines with circles to same-colored lines with crosses). Note that not only is the skill of the original model hindcasts improved, but the skill of the related downscaled and hybrid model hindcasts are as well. Interestingly, the results at Charleston are not as consistent, even though the sea level trend is larger at Charleston than San Diego. This may be due to the larger vertical land motion component at San Diego (Zervas et al., 2013), which the delta-correction could also capture.

Note that an error is introduced for the delta estimate in Month 1 since it uses two separate model hindcasts initialized at two different times. This error could then propagate through the delta-corrected hindcasts for all lead times. The delta-correction degrades the skill of both SEAS5 and the LIM (not shown), whose initializations include observed sea level information that better captures observed trends. Hence, our delta-correction method is only an interim remedy for models with inadequate initialization and/or that do not output the global mean steric component forecasts.

6 Concluding remarks

In this study stemming from the RISE project (a collaboration among scientists at NOAA, NASA/JPL, and several universities), we have considered some key issues in the prediction of coastal SLAs on seasonal time scales, a particularly challenging problem since seasonal forecast systems largely have not been designed with such predictions in mind. Using both deterministic and probabilistic metrics, we assessed the skill of hindcasts from various dynamical and statistical models/techniques — traditional assimilation-initialized seasonal forecast systems based on coupled dynamical models, an empirical regression-based approach (the LIM), and two statistical (linear regression) and dynamical (ECCO adjoint) post-processing techniques applied to output from the seasonal forecast models — against monthly SLAs observed at two sample NOAA NWLON tide gauge stations in San Diego, CA and Charleston, SC. We found that the skill of some of the forecast systems cannot beat a simple “damped persistence” (univariate AR1) approach, especially for Charleston. Even fewer had deterministic or probabilistic skill greater than the LIM (multivariate AR1) approach, which suggests that for future studies, the LIM could serve as a more rigorous benchmark than damped persistence for coastal SLA seasonal forecast skill, both deterministic and probabilistic.

Consistent with previous studies (Long et al. 2021; Shin and Newman 2021), SLA seasonal prediction skill was considerably better for San Diego than Charleston. There are a few possible reasons for poorer Charleston skill. There may simply be lower inherent SLA predictability in the Charleston region. For example, past studies have shown that while ENSO drives a strong and potentially predictable signal in Pacific SLA along the U.S. West coast (Amaya et al. 2022), predictable SLAs along the U.S. East Coast appear associated with Gulf Stream modulation that may have a smaller impact on the predictable monthly signal, compared to unpredictable noise processes (Shin and Newman 2021). This difference is likely exacerbated by large-scale climate model errors in the position of the Gulf Stream, due in part to model grids that are too coarse (e.g., Bryan et al., 2007), although there are also systematic errors in ENSO prediction as well (e.g., Beverley et al. 2023). Current climate model resolution may also be insufficient to entirely capture climate-related signals propagating along the coast, including Kelvin waves driven by ENSO (e.g., Amaya et al. 2022) and other coastally-trapped waves (e.g., Brunner et al. 2019 and references therein; Hughes et al. 2019). Finally, inadequate initialization is also likely a contributing factor in poorly performing forecasts, especially around the Gulf Stream region (Widlansky et al. 2023). For example, the CCSM4 and SPEAR Month 1 hindcasts, which were not initialized using the altimetry observations, failed to correctly represent either the (relatively large) trend along the US East Coast or the interannual component of global mean SLA (Fig. 3). Reliance of the LIM and SEAS5 on ORAS4 and ORAS5, respectively, means that their Charleston initializations are also somewhat deficient relative to San Diego (Fig. 2); since these reanalyses often give little-to-no weighting to satellite altimetry data near the coast (e.g., Balmaseda et al. 2013; Feng et al. 2023), this difference may be related to the fact that while the San Diego tide gauge SLA is correlated with a large scale North Pacific SLA pattern, Charleston is primarily correlated with SLA along the South Atlantic coast (Long et al. 2023). The resolution issues extend to the verification process, where coarse-grid model hindcasts (output at 1° grid resolution) are compared against point observations, rather than against similarly gridded observational datasets. Interestingly, the ECCO adjoint approach, which is "initialized" using a 12-month dynamical spin-up forced by surface observations, has much better Charleston skill compared to the climate models even for Month 1 hindcasts, with significant skill (even after detrending) up to about Month 6 during late spring and late fall. Further diagnosis of these issues, analyzing how inadequate initialization and model error interact with each other so that forecast systems may not take full advantage of sources of predictability, will also need to consider other US tide gauges as well.

A common problem for all the hindcasts is that they do not appear to generate enough categorical "hits", even for categories with the lowest probability thresholds. This issue is evident in individual hindcast ensembles (e.g., Fig. 4), as the ensemble spread is often insufficient relative to observed variability so the model forecasts are not entirely reliable. Note that while we do not expect climate models to predict observed SLA evolution with deterministic certainty, we do expect them to be able to produce ensemble members that can encompass what is observed to occur, so this over-confidence of the model hindcasts is likely also reflective of model error. While both post-processing approaches (downscaling and ECCO adjoint) yield some improvement for deterministic skill relative to the original hindcasts upon which they were based (albeit not uniformly across the models), neither technique appears to improve probabilistic skill and, in some cases, may reduce skill by collapsing the ensemble spread. This suggests the importance of developing post-processing and downscaling methods, whether dynamical or machine learning-based, whose ensembles can capture variability more realistically on the local scales of interest, even if much of that variability is unpredictable.

In this study, we detrended observations and hindcasts using either a linear or quadratic fit to explore the externally-forced trend's potential impact on hindcast skill, including whether different trend estimates could yield different impacts on skill. For Charleston, we found that the pronounced sea level trend increases apparent seasonal prediction skill, especially for hindcasts that are realistically initialized. Much of the hindcast reliability at Charleston also appears to correspond to a trend component. While a trend-related impact on skill also exists for San Diego, it is much less pronounced and complicated by vertical land motion over the length of the hindcast period. For both stations, detrended hindcasts generally remain more skillful than the corresponding damped persistence benchmarks, since the trend also increases apparent persistence. Note that, for the most part, we determined only the qualitative impact of the trend, since these impacts changed when the trend was assumed to be linear or quadratic. A quantitative analysis would require estimating and removing the evolving externally-forced trend, which is complicated by the presence of internal climate variability, vertical land motion, and the apparent trend acceleration since 2011. Of course, if the assumed trend differs from the actual trend, then incorrect detrending could remove some potentially predictable component of natural seasonal-to-interannual variability. This would impact our skill estimates, including metrics of the model's ability to capture observed marginal and conditional probability distributions (e.g., Xu et al. 2022). We also found that removal of the trend can sometimes improve skill for hindcasts with an erroneous trend component, due to inadequate initialization of a sea surface height forecast variable that does not represent the total sea level. Our attempt to correct this issue, by using forecast output to predict month-to-month SLA changes rather than the monthly SLA values themselves, still suffers from an inability to cleanly separate the initial observed state into its trend and seasonal anomaly components, and is thus, at best, a temporary, ad hoc fix. In essence, hindcast skill assessment of U.S. coastal seasonal SLA prediction is also an externally-forced trend detection problem.

This paper represents a multipronged assessment of the skill for seasonal prediction of regional sea level anomalies, involving dynamical, statistical, and hybrid methods. We have tried to provide important information to the climate prediction community about the relative strengths and limitations of various approaches, highlighting the challenges of sea level prediction at sample U.S. coastal stations, and stressing important issues to consider when assessing and comparing sea level prediction skill. Our primary conclusion is that, for the most part, the current seasonal forecasting systems may not yet be fit for the purpose of making coastal sea level predictions in the regions considered here. It is apparent that making useful predictions of coastal SLA is a hard test for seasonal forecast systems, which helps identify needs for additional improvement in both climate models and their initialization. To make progress, we propose studies aimed at the following:

- Evaluating the extent to which higher model resolution, which could reduce large scale model errors such as Gulf Stream position and strength, will improve forecasts that also depend upon complex coastal geography and bathymetry.
- Developing non-Boussinesq ocean models that include the global ocean volume changes, both barystatic and thermosteric, which are important to local sea level

prediction, or alternatively developing models whose output include global changes in steric volume as an additional diagnostic.

- Understanding how best to initialize climate models (either Boussinesq or non-Boussinesq) so that their coastal forecasts may be best used, either directly or through post-processing and downscaling.
- Investigating methods to improve the reliability of climate model forecast ensemble spread (e.g., stochastic parameterization; Sardeshmukh et al. 2023).
- Diagnosing drivers of model error in hindcast ensembles, especially the rapid initial development of error (including the initialization drift of the mean dynamic topography of the ocean) that degrades SLA forecast skill.

Additionally, the issues discussed concerning how to post-process/downscale seasonal climate forecasts lead us to suggest studies focused on:

- Evaluating how climate model forecast ensembles may be used for driving smaller-scale, limited domain ocean models, with better coastal processes including tide, wave, and ocean dynamic effects, that are all necessary for providing actionable coastal information.
- Constructing forecast ensembles that can capture prediction uncertainty both at the large climate scales and at the smaller atmospheric and oceanic scales relevant to the coastal regions.
- Evaluating new empirical and machine learning approaches, designed both to post-process model forecasts and to make coastal SLA seasonal predictions outright, as alternative solutions while the suggestions above are evaluated.

Acknowledgments.

XL, MN, and SS acknowledge the support from NOAA cooperative agreements NA17OAR4320101 and NA22OAR4320151, and U.S. DoC/ NOAA/Bipartisan Infrastructure Law (BIL). MJW was supported by the NOAA Climate Program Office's Modeling, Analysis, Predictions, and Projections (MAPP) program through grant NA22OAR4310138. Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004).

Data Availability Statement.

All hindcast and observational datasets discussed in this paper are available for viewing and downloading at https://www.psl.noaa.gov/forecasts/SeaLevel/#RISE.

Table 1 Key characteristics of the ocean model component in the three dynamical forecast systems, including the name of the forecast system, ensemble size of the seasonal hindcasts, lead time (months), the name of the ocean model component, nominal horizontal resolution, whether the system includes a global mean sea surface height component in its output, whether the system assimilates the altimetry observed sea surface height into its initial conditions, and the reference related to each of the seasonal forecast systems.

Name	Ensemble Size	Lead Time	Ocean Model	Grid Resolution (deg)	Global Mean Sea Level	Altimetry-initialized?	References
CCSM4	10	11	POP2	1	No	No	Kirtman et al. (2014)
SPEAR	15	11	MOM6	0.5	Partly	No	Delworth et al. (2020)
SEAS5	25	6	NEMO	0.25	Yes	Yes	Johnson et al. (2019)

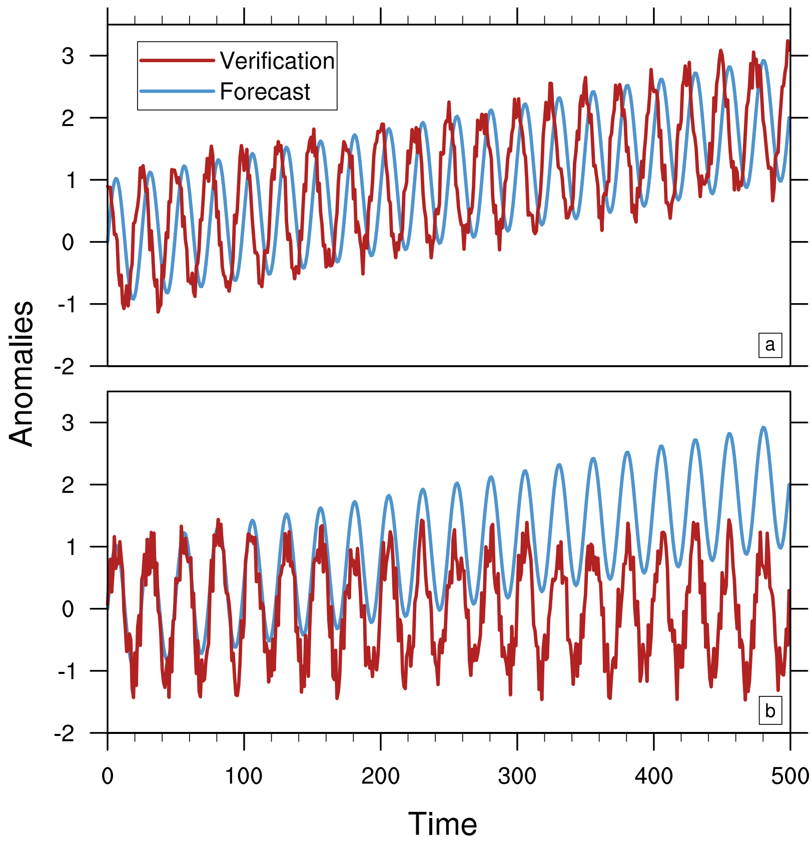


Figure 1. (a) An example of how superposing two uncorrelated time series onto a linear trend generates the new well-correlated time series. The original uncorrelated time series has a standard deviation of 0.9, and the superposed linear trend has a standard deviation of 0.56. The two new time series correlate 0.39. (b) Superposing different trends can reduce the correlation of two otherwise correlated time series from 0.90 to 0.67.

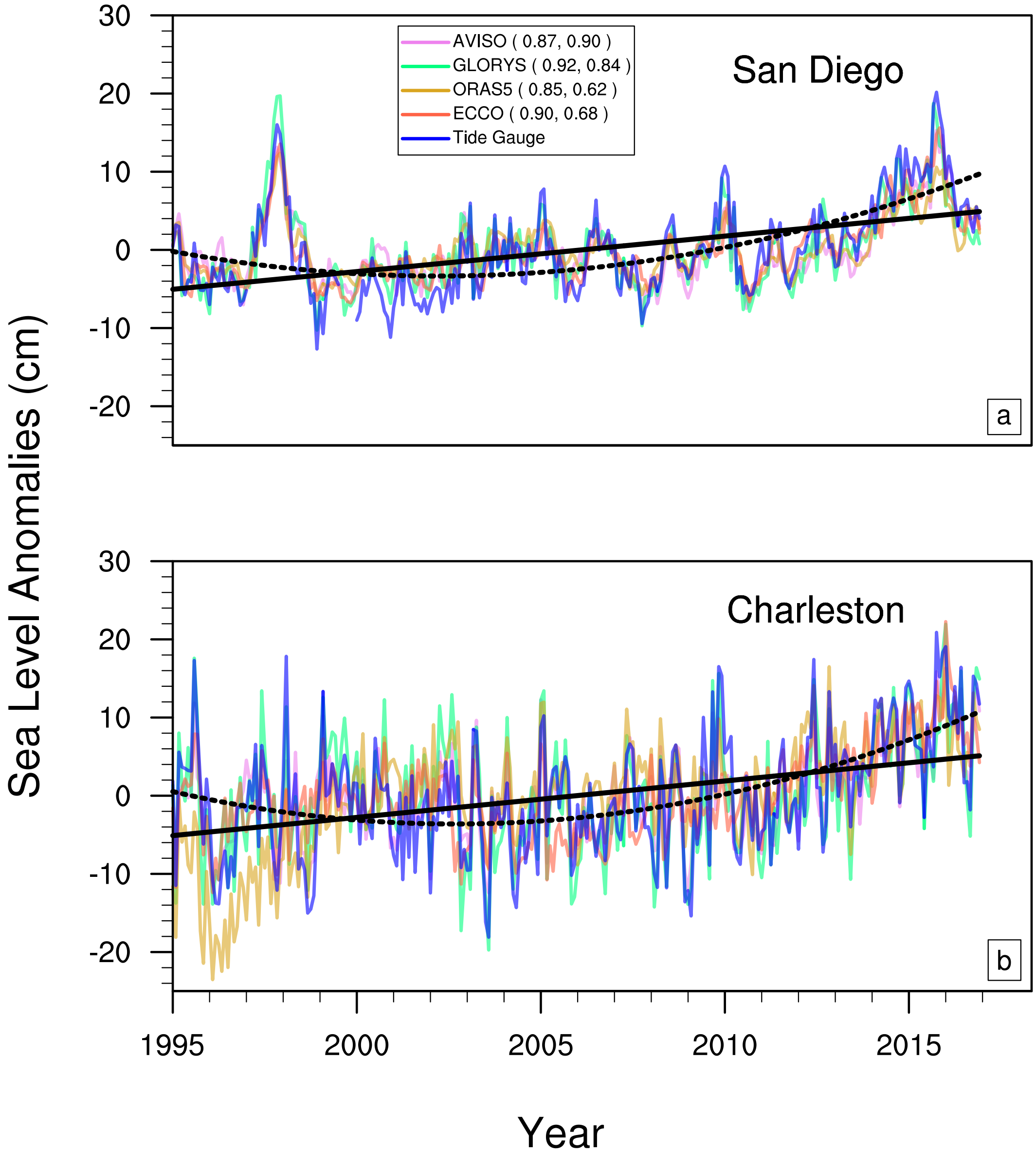


Figure 2. Monthly mean SLA from two tide gauge stations, (top) San Diego and (bottom) Charleston, compared with SLA values of the nearest grid point from the observationally-based reanalysis datasets, AVISO, GLORYS12, ECCO, and ORAS5. The correlation coefficient between each reanalysis and the corresponding tide gauge time series is given as numbers in parenthesis (San Diego, Charleston). Linear and quadratic least square fits to each tide gauge time series are also shown.

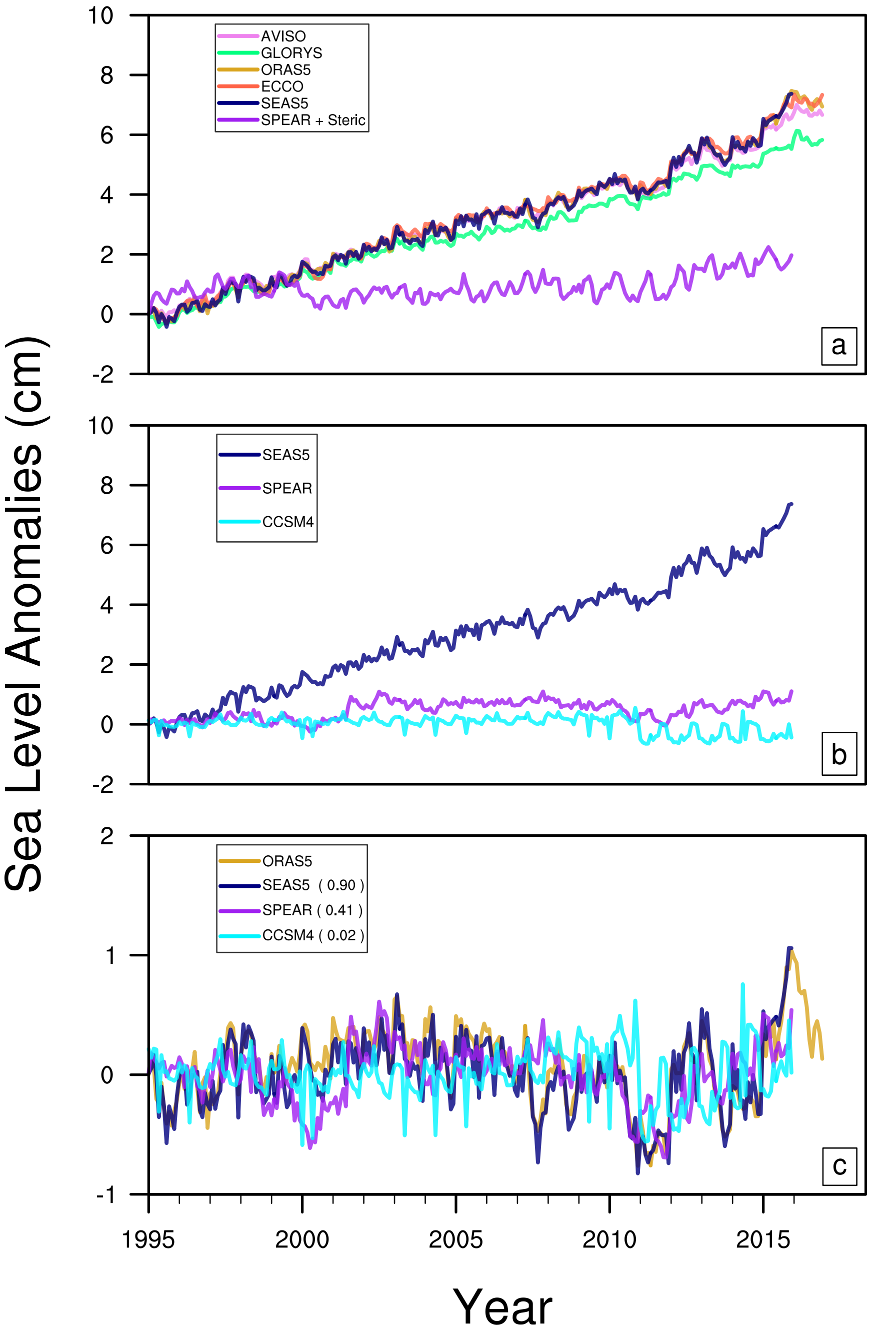


Figure 3. Global monthly mean SLA from satellite observations (AVISO) and observationally-based reanalyses (GLORYS12, ECCO, and ORAS5) compared to the Month 1 hindcast from SEAS5, SPEAR, and CCSM4. The anomalies are relative to the climatology of 1995-2016 and relative to the global mean in January 1995. (a) Comparison of observed values with SEAS5 and SPEAR hindcasts, where the latter is corrected with its global mean steric component determined from the temperature and salinity profiles. (b) Comparison of Month 1 values from the SEAS5 and original (without global mean steric correction) SPEAR and CCSM4 hindcasts. (c) Comparison of linearly detrended ORAS5 with linearly trended

Month 1 values from the SEAS5, SPEAR, and CCSM4 hindcasts; the number in the legend indicates the correlation of each hindcast time series with ORAS5.

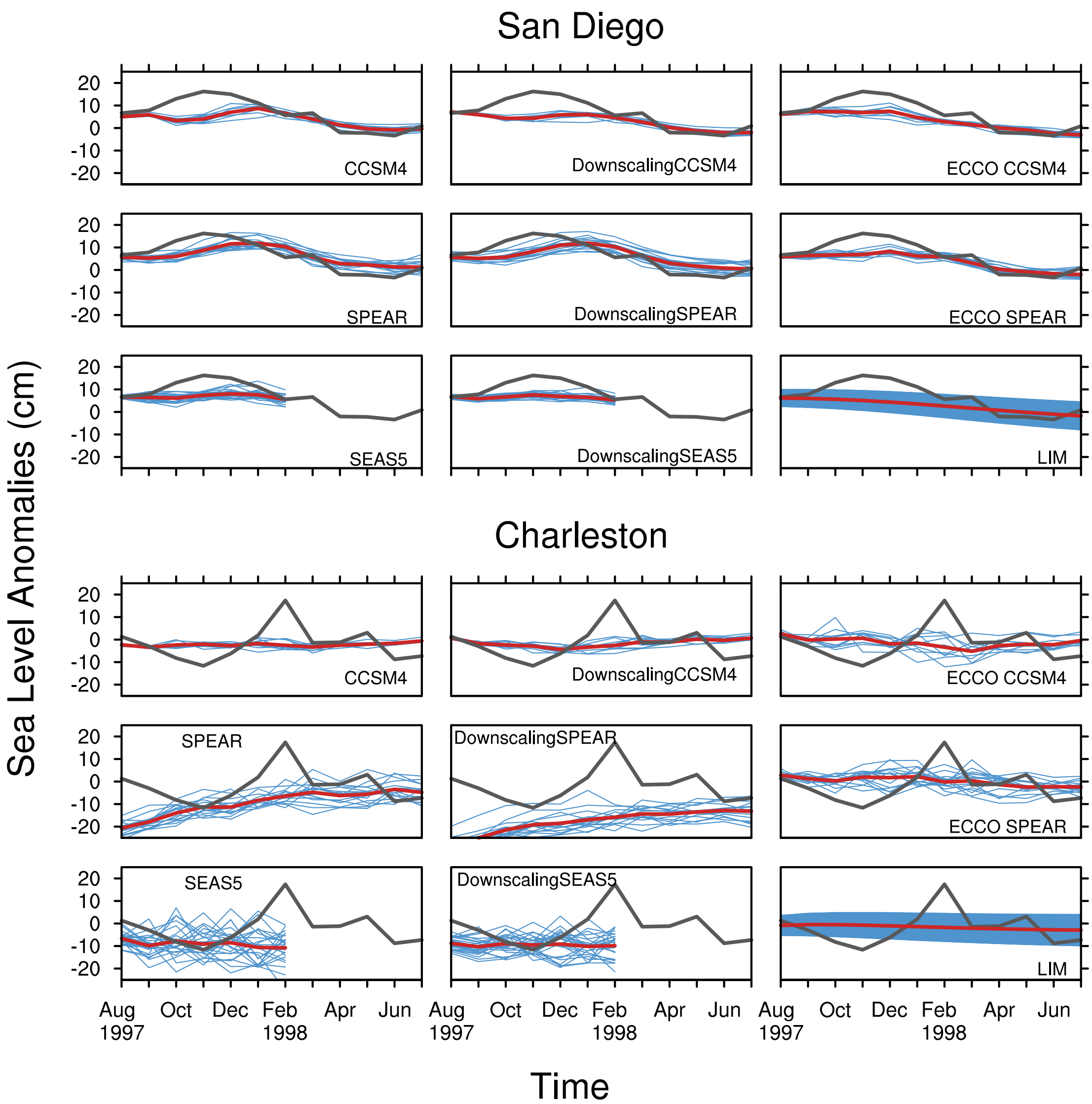


Figure 4. Observed (gray) and predicted (red and blue) monthly SLA anomalies from August 1997 to July 1998 at San Diego (top) and Charleston (bottom). Observations (dark gray) are from the tide gauge records, and the model hindcasts were initialized by August 1, 1997 (LIM was initialized in July 1997). The red solid line is the ensemble mean forecast, and the blue solid lines/ shading indicates ensemble members/spread. Units are cm.

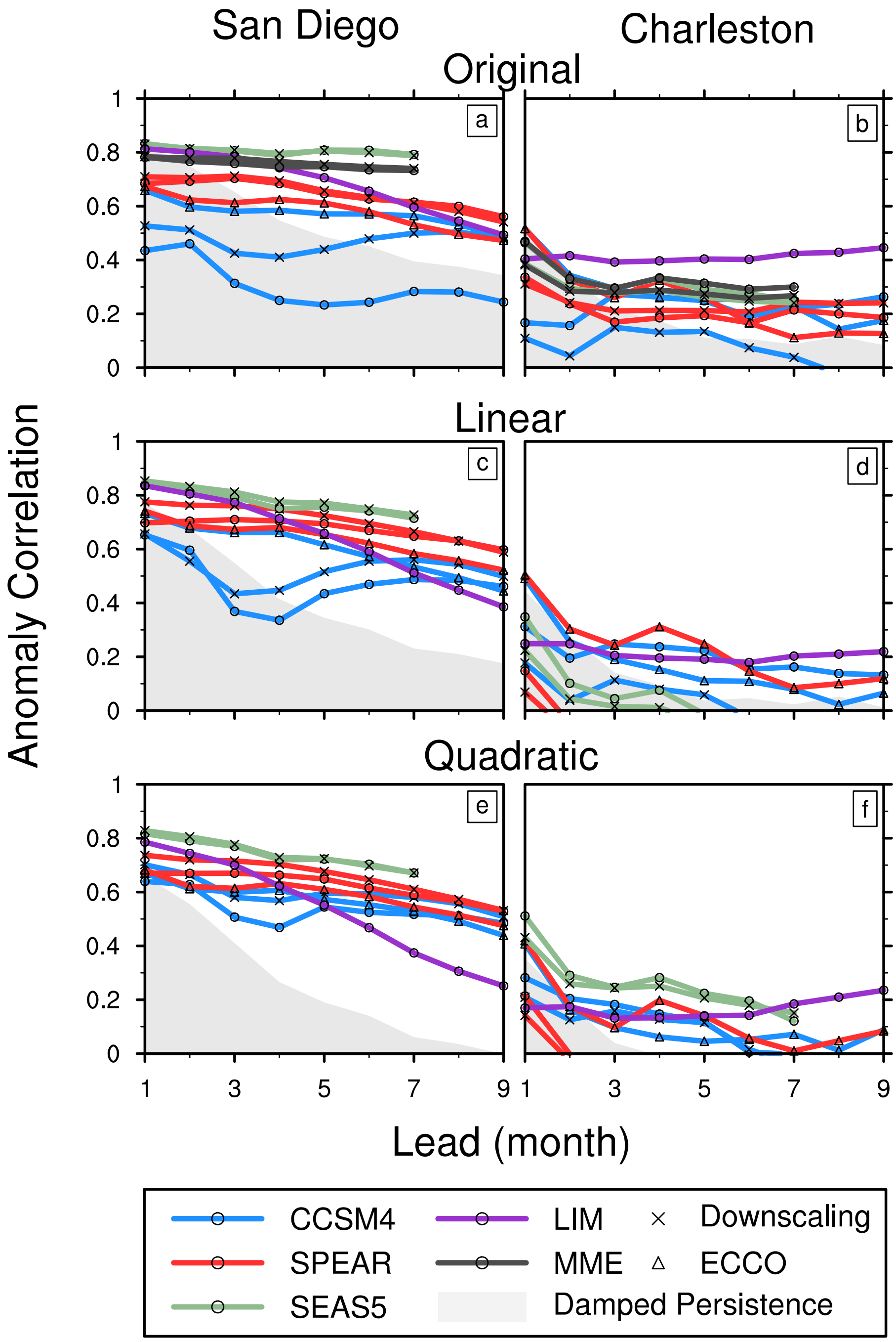


Figure 5. Deterministic skill measured by anomaly correlation coefficient (ACC) between the hindcasts and tide gauge observations at (left) San Diego and (right) Charleston at different lead times. The verification period is from 1995 to 2015, using hindcasts initialized in all calendar months. Gray shading shows damped persistence skill. Top row: Skill of each forecast technique for (a) San Diego and (b) Charleston. Second row: Same as (a and b) but after linear detrending of the observed tide gauge time series and from each hindcast. Third row: Same as (c and d) but using quadratic detrending. Note that trending also impacts the damped persistence time scale and skill.

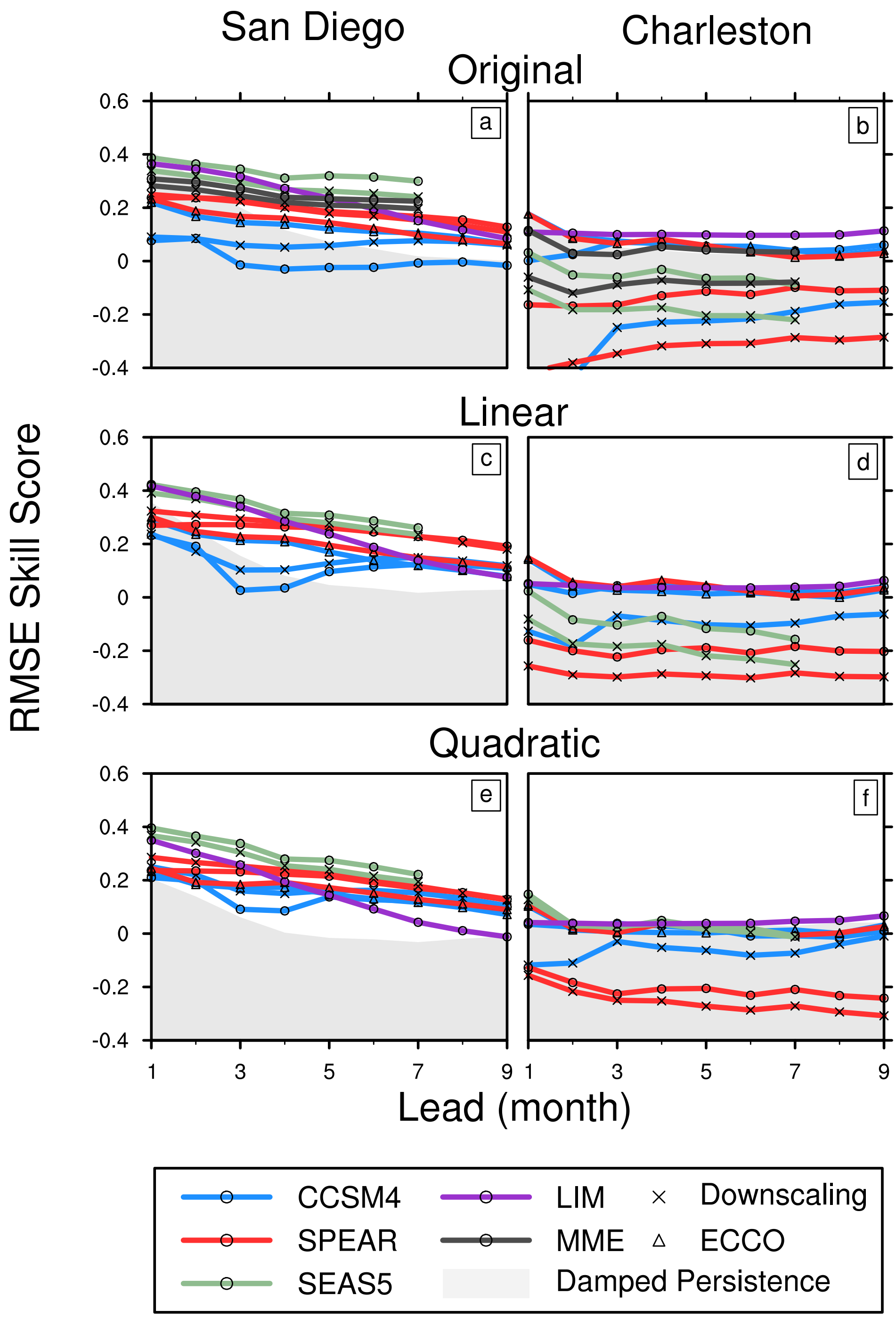


Figure 6. Same as Fig. 5 but using RMS skill score (RMSSS).

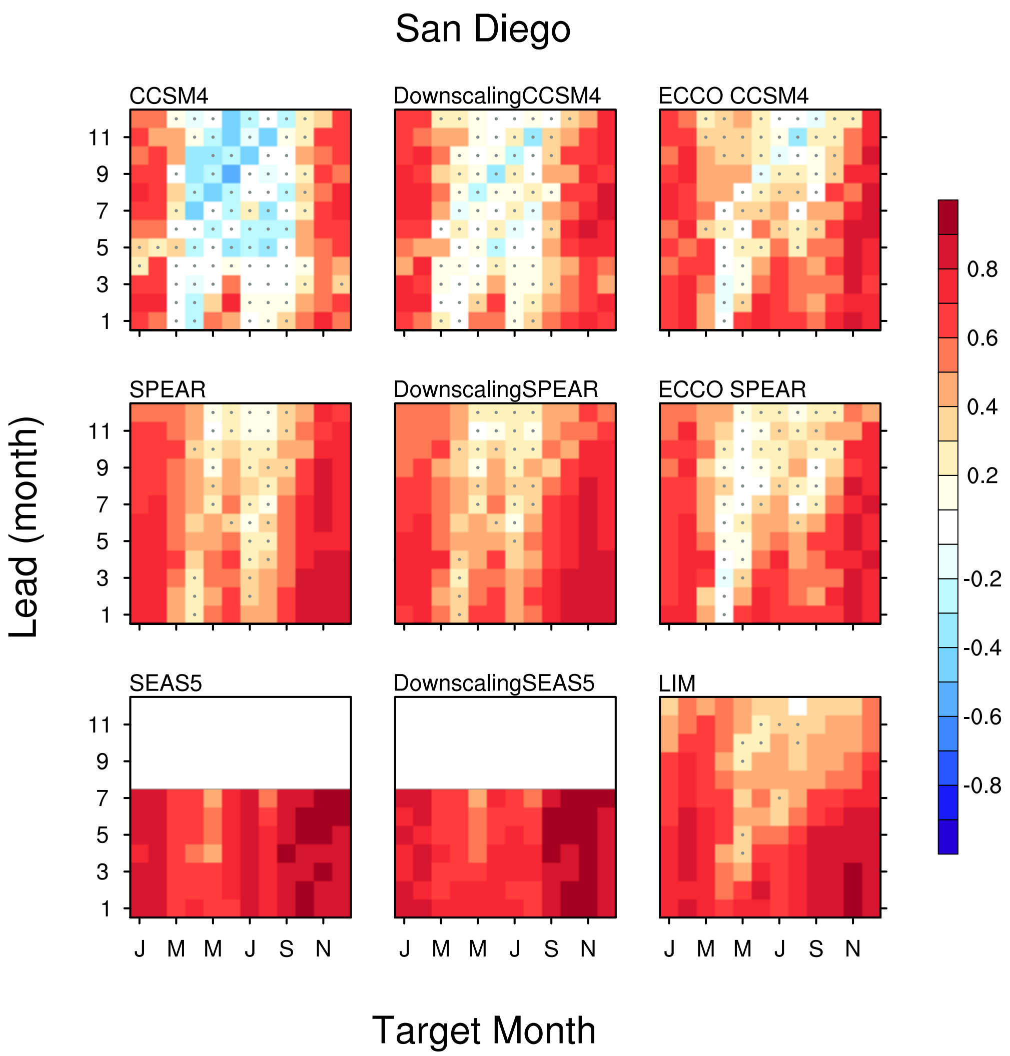
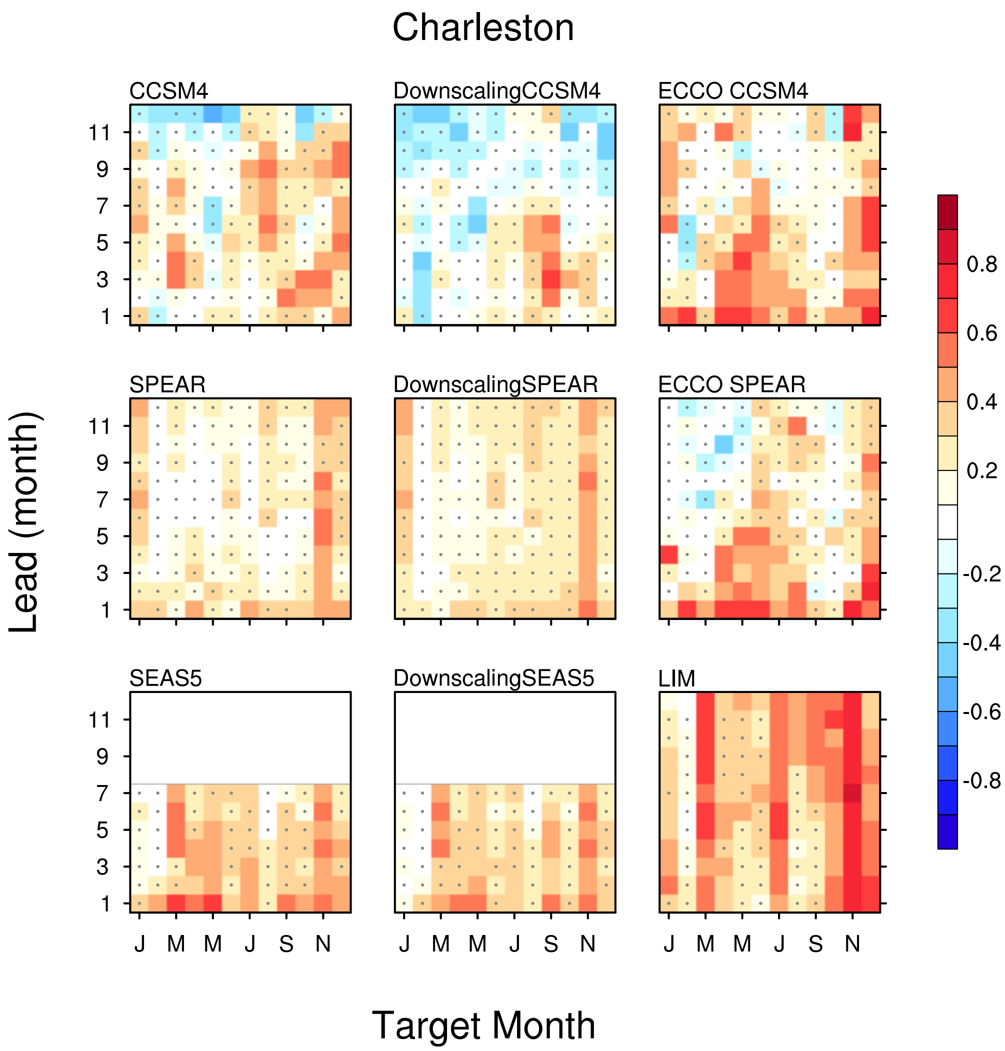


Figure 7. Deterministic skill measured by anomaly correlation coefficient (ACC) between the hindcasts and tidal gauge observations at San Diego at different lead times and target months. The gray dots indicate anomaly correlation values that are not significant at the 0.1 level using a two-tail student t-test. The verification period is 1995 to 2015. No detrending is performed upon either hindcasts or observations.



867
868 Figure 8. Same as Fig. 7 but for the Charleston tide gauge.

869

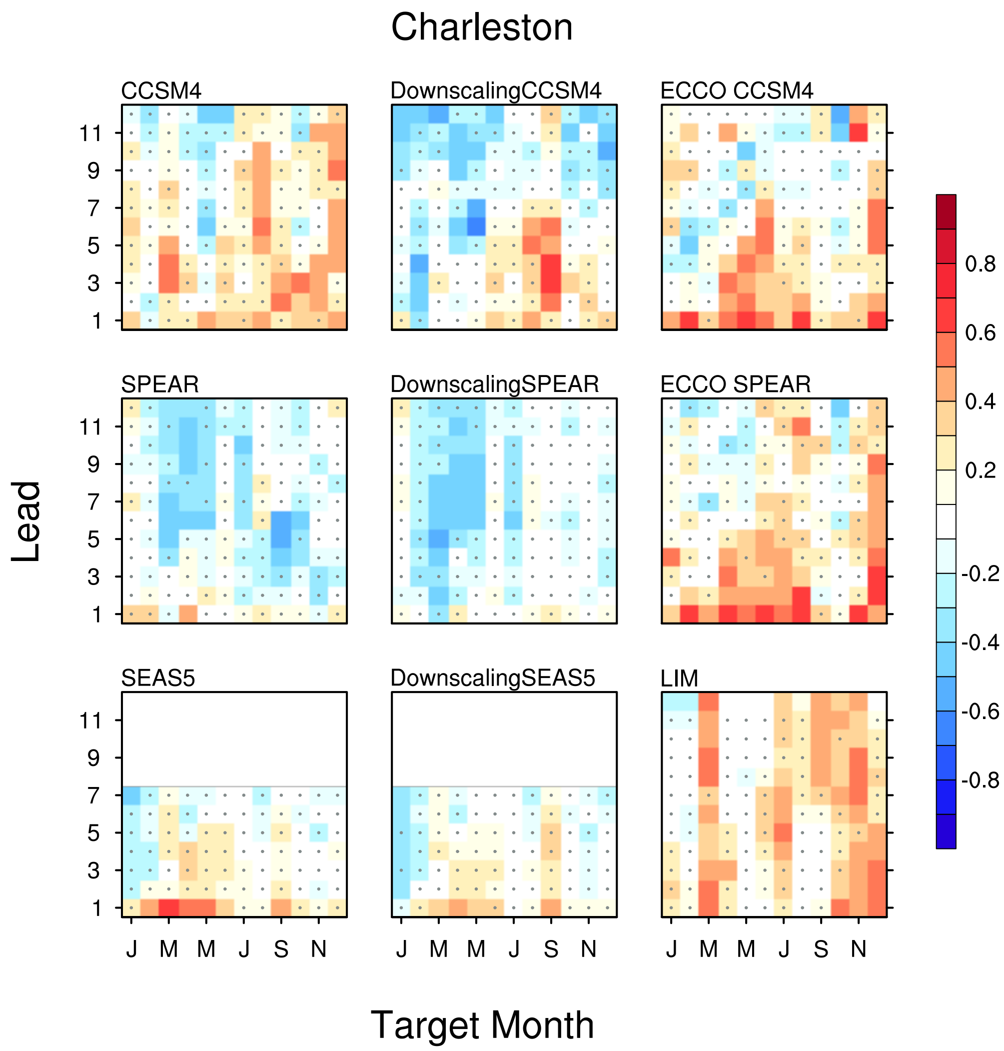


Figure 9. The same as Fig. 8 but computed after linearly detrending observations and hindcasts.

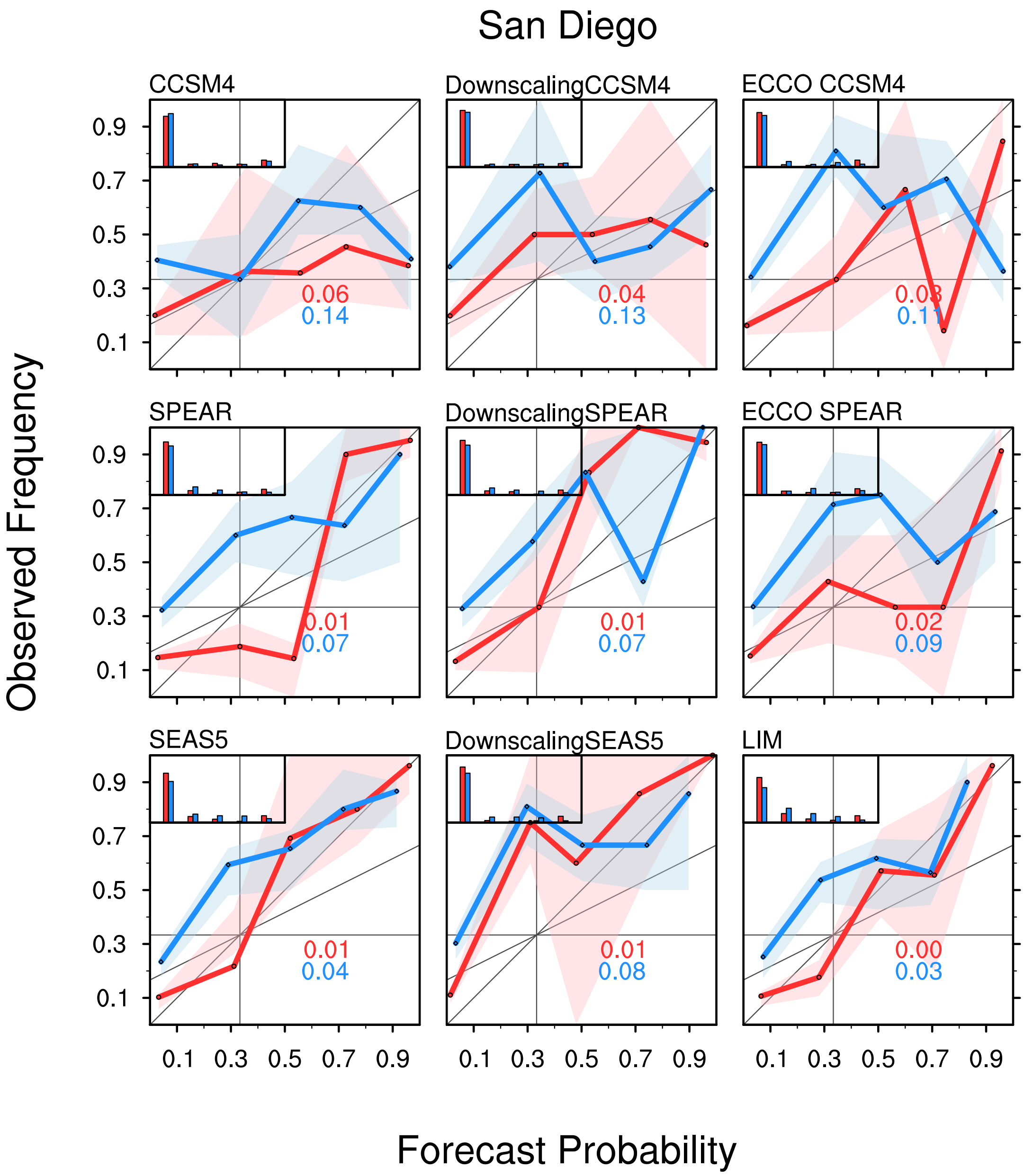
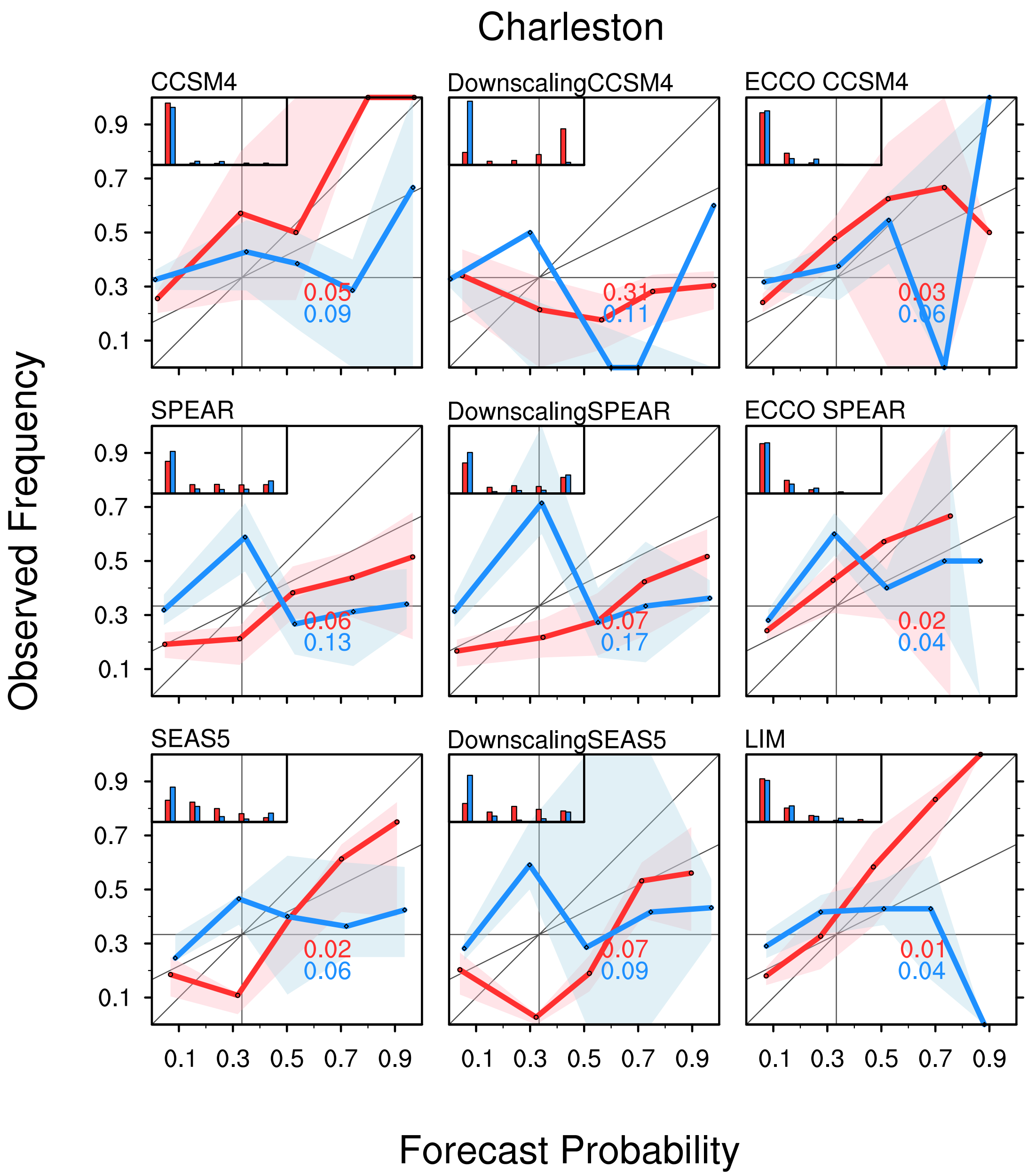


Figure 10. Reliability and sharpness diagrams of each hindcast for San Diego at Month 4. The mean forecast probability is plotted against the mean observed frequency for the reliability curve, determined by averaging all hindcasts within each quintile bin category. Red is for upper tercile hindcast, and blue is for lower tercile hindcast. The annotation is the reliability value with the same color coding (note that lower values represent better reliability). Gray shading shows the uncertainty of the reliability curves based on a bootstrapping calculation.



883

884 Figure 11. Same as Fig. 10 but for Charleston.

885

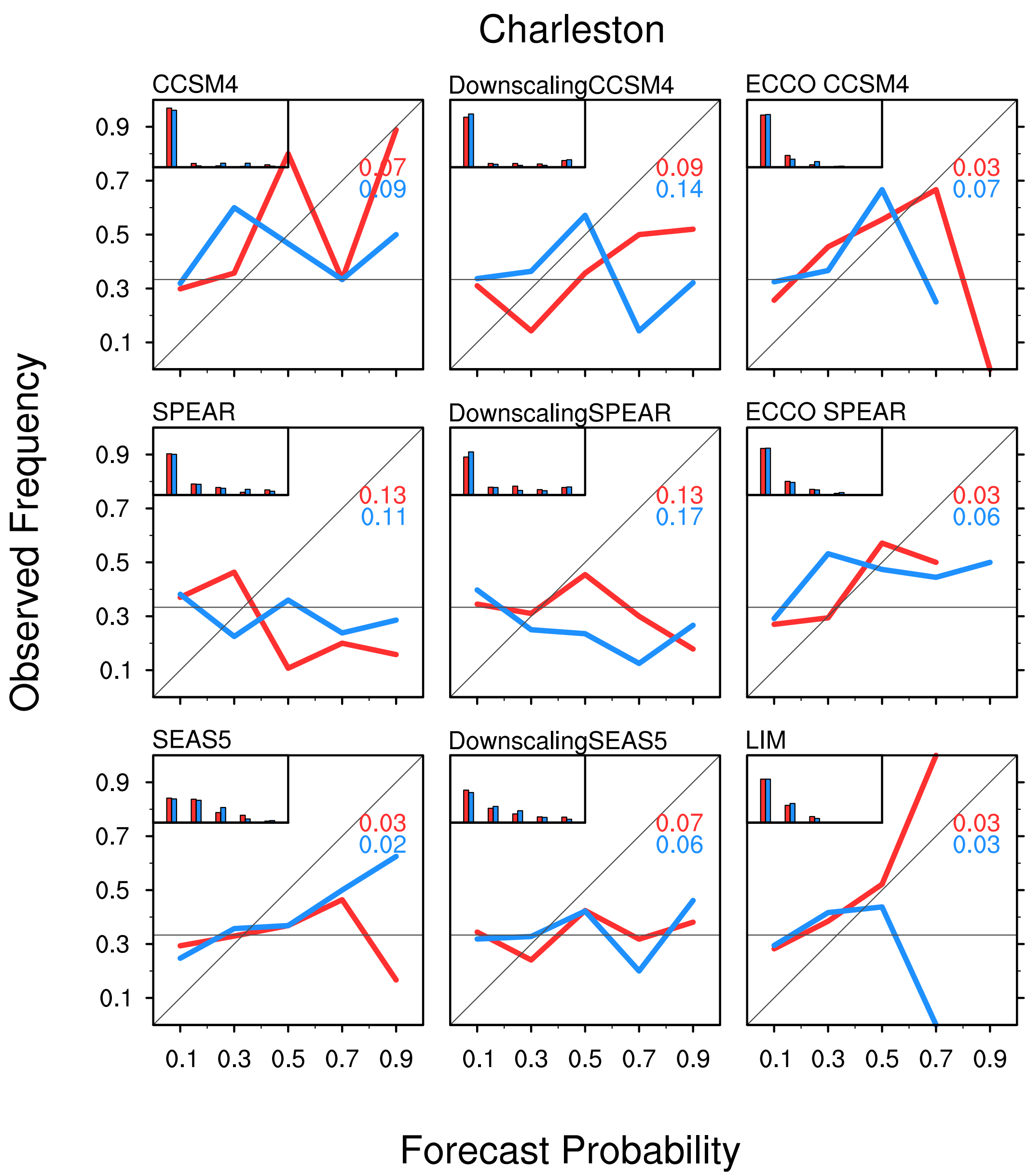


Figure 12. The same as Fig. 11 but for reliability and sharpness calculated after the linear trend is removed from the observed tide gauge time series and each of the hindcasts. The terciles of the observations are also calculated using the detrended data.

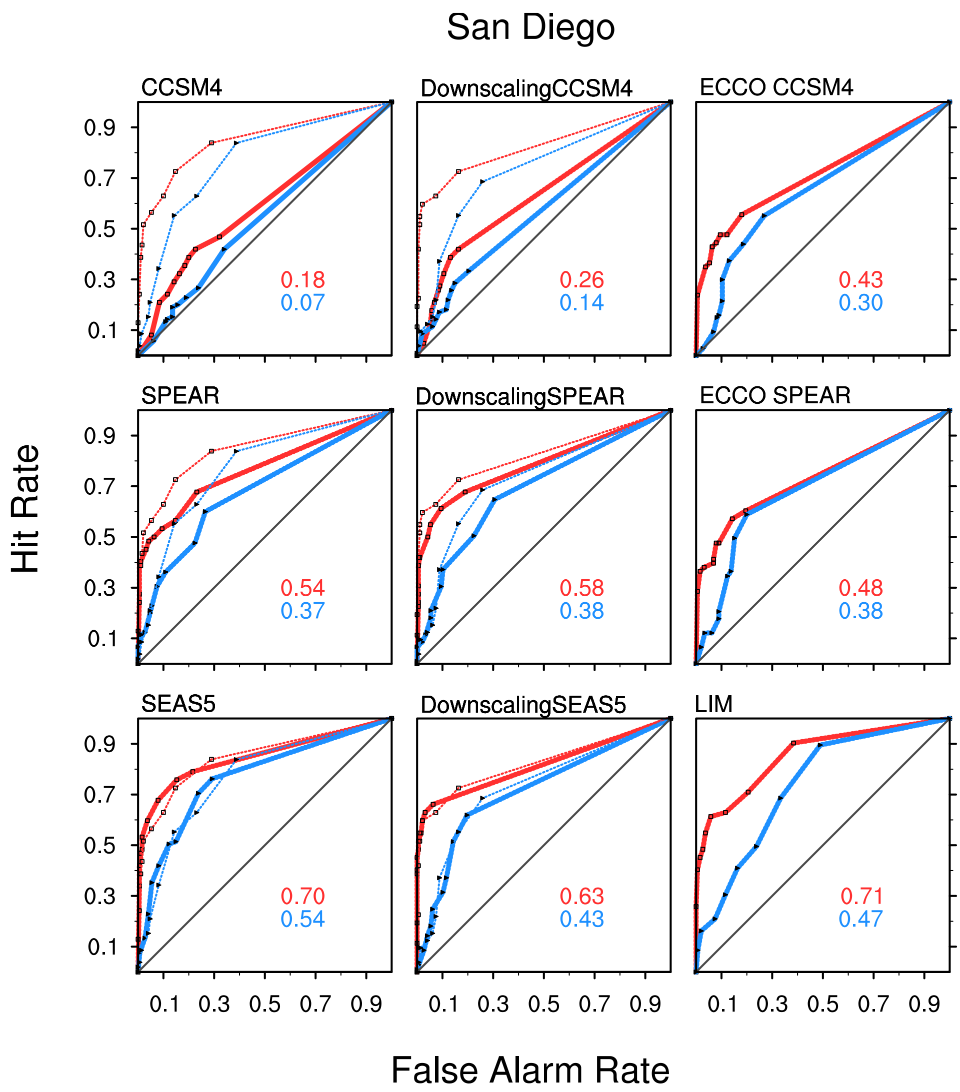


Figure 13. ROC curve of each of the hindcasts for San Diego at Month 4. Red is for the upper tercile forecast, and blue is for the lower tercile forecast. The ROC skill score (ROCS) is shown in each panel. The dashed lines in the first and second columns are the ROC curves for the multi-model mean of the hindcasts from three dynamical models and three downscaled versions, respectively.

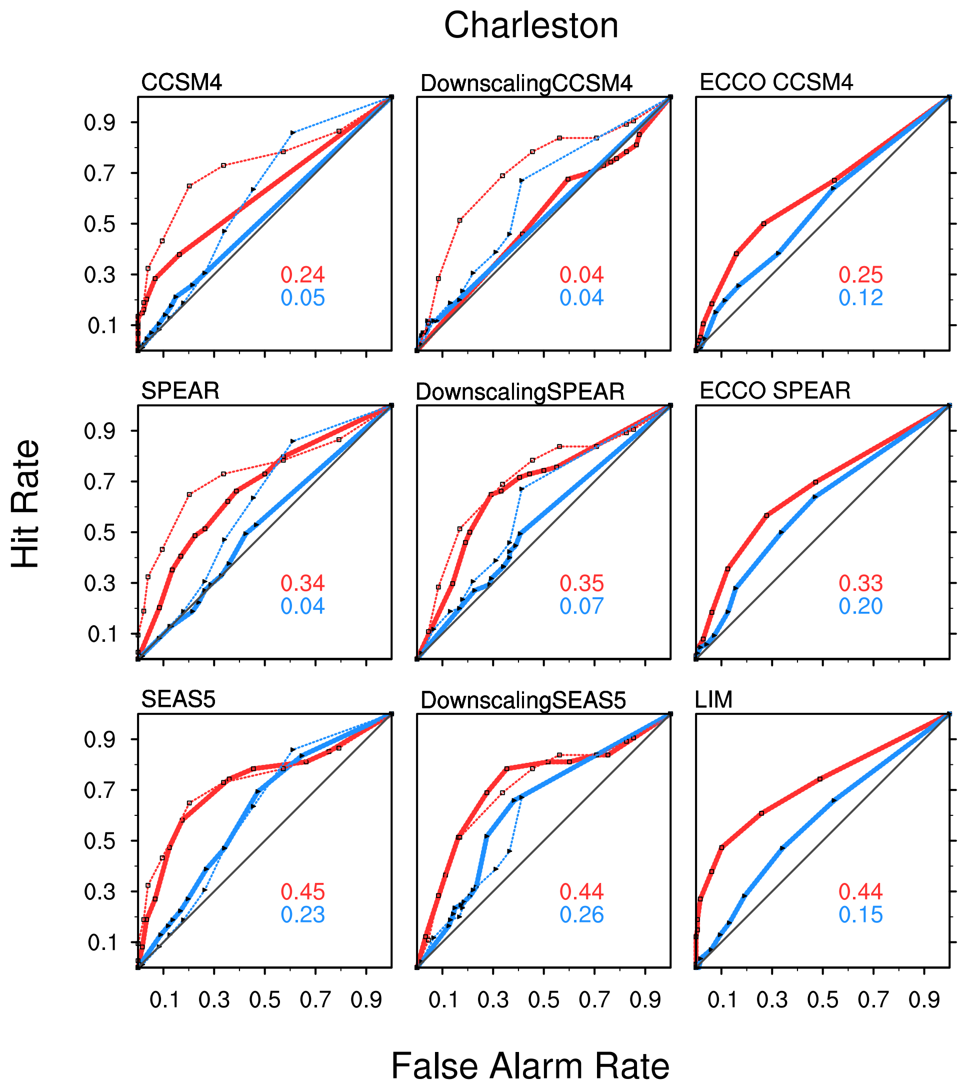


Figure 14. Same as Fig. 13 but for Charleston.

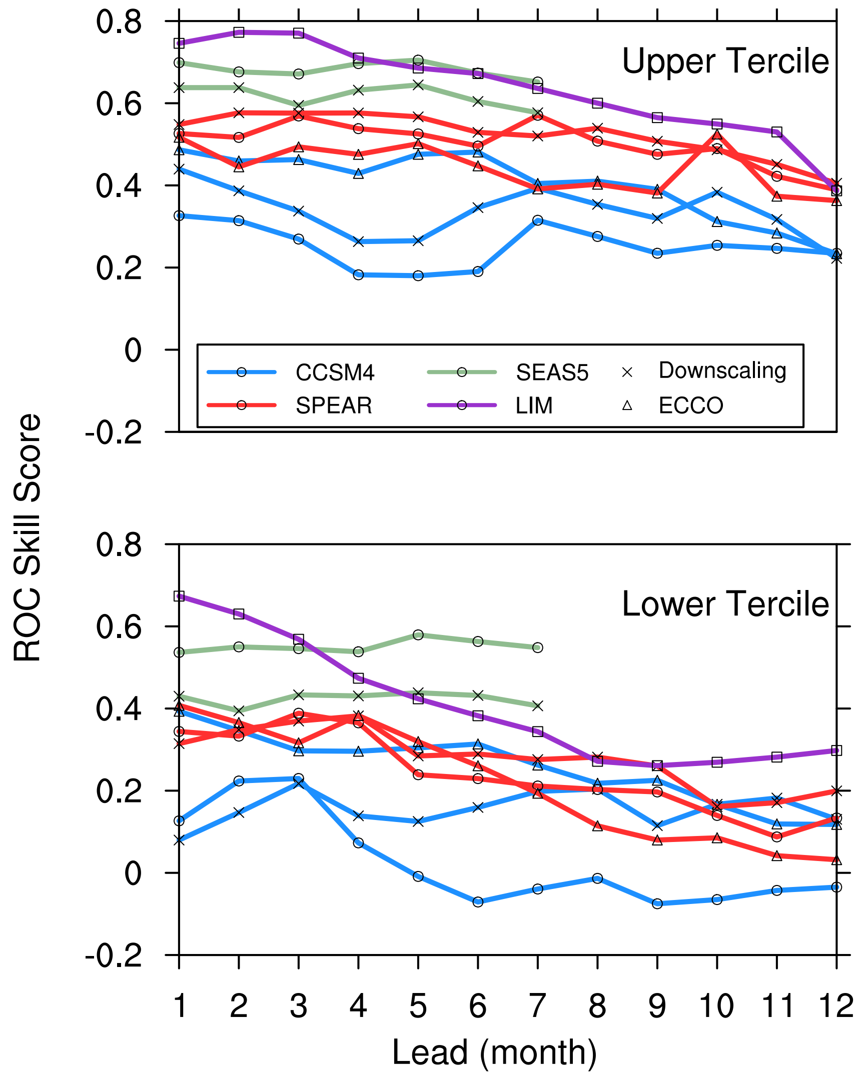


Figure 15. The ROC skill score (ROCS) for upper and lower tercile hindcast for each prediction technique for San Diego at different lead times.

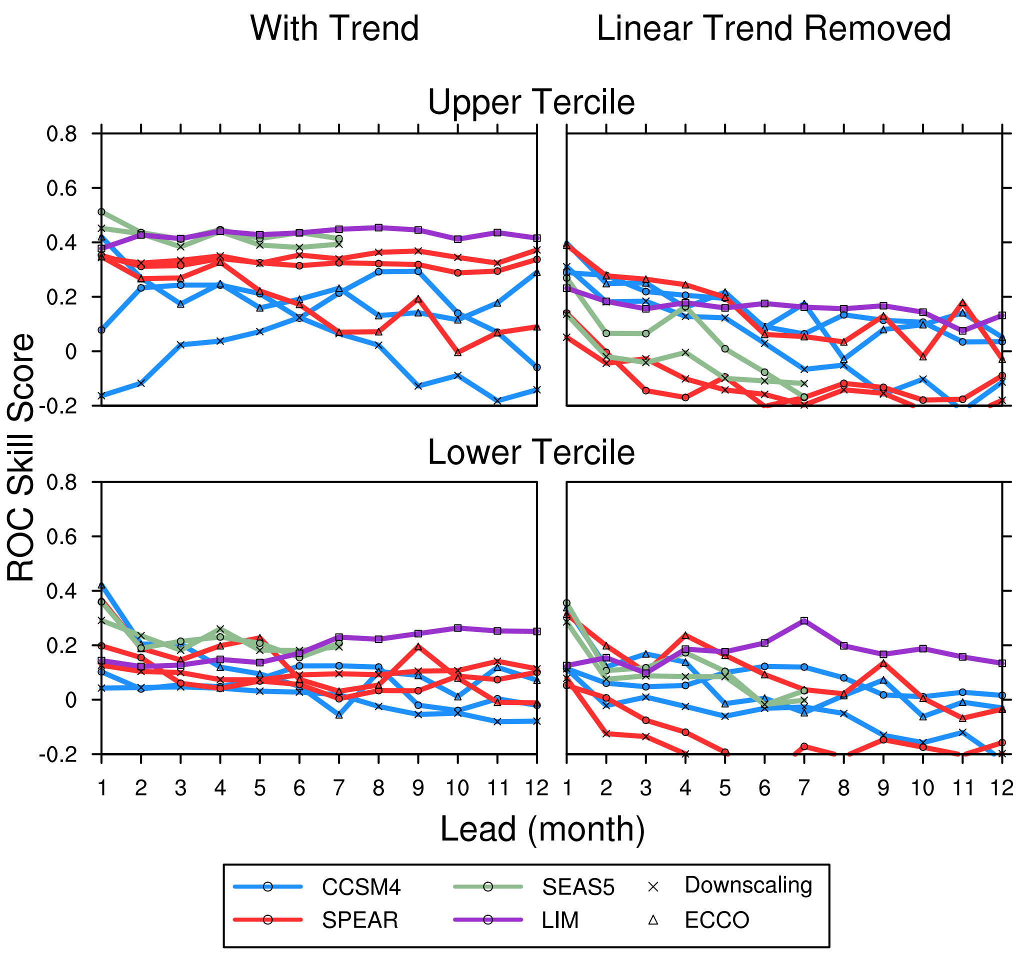


Figure 16. The ROC skill score (ROCS) for upper and lower tercile hindcast for each of the prediction techniques, verified for Charleston at different lead times, determined for (left column) full fields and (right column) linearly detrended fields.

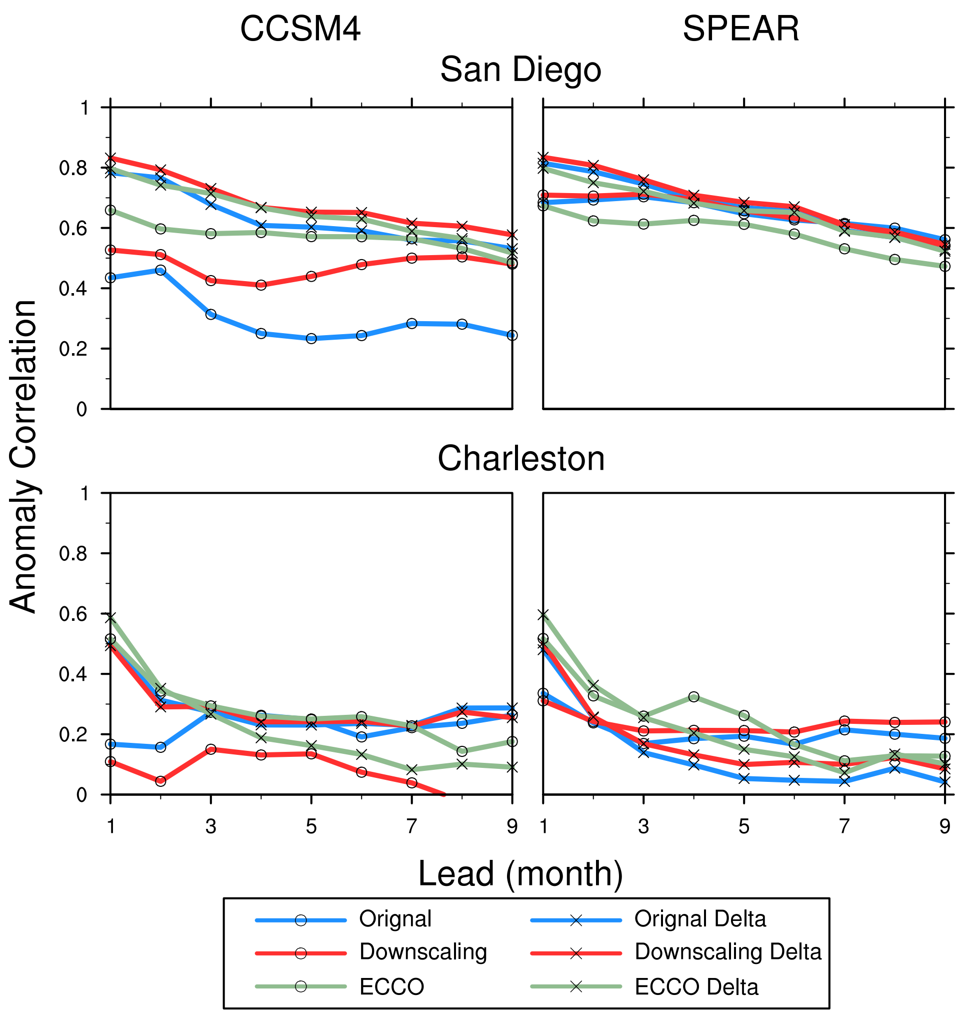


Figure 17. Anomaly correlation coefficient (ACC) between the hindcasts and tidal gauge observations at San Diego and Charleston at different lead times. For each of the model hindcasts, the ACC of the delta corrected hindcast is compared with the skill of the original hindcast. See text for details of the procedure of creating the delta corrected hindcast.

References

Amaya DJ, MG Jacox, J Dias, MA Alexander, KB Karnauskas, JD Scott, and M Gehne (2022). Subseasonal-to-seasonal forecast skill in the California Current System and its connection to coastal Kelvin waves. *JGR: Oceans*. https://doi.org/10.1029/2021JC017892.

Balmaseda, M. A., Mogensen, K., & Weaver, A. T. (2013). Evaluation of the ECMWF ocean reanalysis system ORAS4. *Quarterly Journal of the Royal Meteorological Society*, *139*(674), 1132-1161.

Balmaseda, M. A., Vitart, F., Ferranti, L., & Anderson, D. (2002). *Westerly wind events and the 1997 El Nino event in the ECMWF seasonal forecasting system: A case study*. ECMWF.

Beverley, J., Newman, M., & Hoell, A. (2023). Rapid Development of Systematic ENSO-Related Seasonal Forecast Errors. *Geophysical Research Letters*, *50*(10), e2022GL102249.

Brunner, K., D. Rivas, and K. M. M. Lwiza, 2019: Application of Classical Coastal Trapped Wave Theory to High-Scattering Regions. *J. Phys. Oceanogr.*, **49**, 2201–2216

Bryan, F. O., Hecht, M. W., & Smith, R. D. (2007). Resolution convergence and sensitivity studies with North Atlantic circulation models. Part I: The western boundary current system. *Ocean Modelling*, *16*(3), 141-159. https://doi.org/https://doi.org/10.1016/j.ocemod.2006.08.005

Cazenave, A., Henry, O., Munier, S., Delcroix, T., Gordon, A., Meyssignac, B., Llovel, W., Palanisamy, H., & Becker, M. (2012). Estimating ENSO influence on the global mean sea level, 1993–2010. *Marine Geodesy*, *35*(sup1), 82-97.

Chowdhury, M. R., Chu, P. S., Schroeder, T., & Colasacco, N. (2007). Seasonal sea-level forecasts by canonical correlation analysis—An operational scheme for the US-affiliated Pacific Islands. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *27*(10), 1389-1402.

Church, J. A., Clark, P. U., Cazenave, A., Gregory, J. M., Jevrejeva, S., Levermann, A., Merrifield, M. A., Milne, G. A., Nerem, R. S., & Nunn, P. D. (2013). Sea-level rise by 2100. *Science*, *342*(6165), 1445-1445.

Church, J. A., & White, N. J. (2006). A 20th century acceleration in global sea-level rise. *Geophysical Research Letters*, *33*(1).

Consortium, E., Fukumori, I., Wang, O., Fenty, I., Forget, G., Heimbach, P., & Ponte, R. (2020). ECCO central estimate (version 4 release 4). In.

Delworth, T. L., Cooke, W. F., Adcroft, A., Bushuk, M., Chen, J. H., Dunne, K. A., Ginoux, P., Gudgel, R., Hallberg, R. W., & Harris, L. (2020). SPEAR: The next generation GFDL modeling system for seasonal to multidecadal prediction and projection. *Journal of Advances in Modeling Earth Systems*, *12*(3), e2019MS001895.

Deser, C., Phillips, A. S., Alexander, M. A., & Smoliak, B. V. (2014). Projecting North American climate over the next 50 years: Uncertainty due to internal variability. *Journal of Climate*, *27*(6), 2271-2296.

Ding, H., Newman, M., Alexander, M. A., & Wittenberg, A. T. (2018). Skillful climate forecasts of the tropical Indo-Pacific Ocean using model-analogs. *Journal of Climate*, *31*(14), 5437-5459.

Dusek, G., Sweet, W. V., Widlansky, M. J., Thompson, P. R., & Marra, J. J. (2022). A novel statistical approach to predict seasonal high tide flooding [Original Research]. *Frontiers in Marine Science*, *9*. https://doi.org/10.3389/fmars.2022.1073792

Forget, G., Campin, J.-M., Heimbach, P., Hill, C., Ponte, R., & Wunsch, C. (2015). ECCO version 4: An integrated framework for non-linear inverse modeling and global ocean state estimation. *Geoscientific Model Development*, *8*(10), 3071-3104.
Frankignoul, C., Gastineau, G., & Kwon, Y.-O. (2017). Estimation of the SST response to anthropogenic and external forcing and its impact on the Atlantic multidecadal oscillation and the Pacific decadal oscillation. *Journal of Climate*, *30*(24), 9871-9895.
Frederikse, T., Lee, T., Wang, O., Kirtman, B., Becker, E., Hamlington, B., Limonadi, D., & Waliser, D. (2022). A Hybrid Dynamical Approach for Seasonal Prediction of Sea-Level Anomalies: A Pilot Study for Charleston, South Carolina. *Journal of Geophysical Research: Oceans*, *127*(8), e2021JC018137. https://doi.org/https://doi.org/10.1029/2021JC018137
Garcia-Soto, C., Cheng, L., Caesar, L., Schmidtko, S., Jewett, E. B., Cheripka, A., Rigor, I., Caballero, A., Chiba, S., & Báez, J. C. (2021). An overview of ocean climate change indicators: Sea surface temperature, ocean heat content, ocean pH, dissolved oxygen concentration, arctic sea ice extent, thickness and volume, sea level and strength of the AMOC (Atlantic Meridional Overturning Circulation). *Frontiers in Marine Science*, *8*, 642372.
Gill, S. K., & Schultz, J. R. (2001). Tidal datums and their applications.
Griffies, S. M., & Adcroft, A. J. (2008). Formulating the equations of ocean models. *Ocean modeling in an eddying regime*, *177*, 281-317.
Griffies, S. M., Danabasoglu, G., Durack, P. J., Adcroft, A. J., Balaji, V., Böning, C. W., Chassignet, E. P., Curchitser, E., Deshayes, J., & Drange, H. (2016). OMIP contribution to CMIP6: Experimental and diagnostic protocol for the physical component of the Ocean Model Intercomparison Project. *Geoscientific Model Development*, 3231.
Griffies, S. M., & Greatbatch, R. J. (2012). Physical processes that impact the evolution of global mean sea level in ocean climate models. *Ocean Modelling*, *51*, 37-72. https://doi.org/https://doi.org/10.1016/j.ocemod.2012.04.003
Hague, B. S., McGregor, S., Jones, D. A., Reef, R., Jakob, D., & Murphy, B. F. (2023). The Global Drivers of Chronic Coastal Flood Hazards Under Sea-Level Rise. *Earth's Future*, *11*(8), e2023EF003784. https://doi.org/https://doi.org/10.1029/2023EF003784
Hamlington, B. D., Gardner, A. S., Ivins, E., Lenaerts, J. T., Reager, J., Trossman, D. S., Zaron, E. D., Adhikari, S., Arendt, A., & Aschwanden, A. (2020). Understanding of contemporary regional sea-level change and the implications for the future. *Reviews of Geophysics*, *58*(3), e2019RG000672.
Hamlington, B. D., Leben, R. R., Kim, K.-Y., Nerem, R. S., Atkinson, L. P., & Thompson, P. R. (2015). The effect of the El Niño-Southern Oscillation on U.S. regional and coastal sea level. *Journal of Geophysical Research: Oceans*, *120*(6), 3970-3986. https://doi.org/https://doi.org/10.1002/2014JC010602
Hamlington, B. D., Piecuch, C. G., Reager, J. T., Chandanpurkar, H., Frederikse, T., Nerem, R. S., Fasullo, J. T., & Cheon, S.-H. (2020). Origin of interannual variability in global mean sea level. *Proceedings of the National Academy of Sciences*, *117*(25), 13983-13990. https://doi.org/doi:10.1073/pnas.1922190117
Han, W., Meehl, G. A., Stammer, D., Hu, A., Hamlington, B., Kenigson, J., Palanisamy, H., & Thompson, P. (2017). Spatial patterns of sea level variability associated with natural internal climate modes. *Integrative study of the mean sea level and its components*, 221-254.
Han, W., Stammer, D., Thompson, P., Ezer, T., Palanisamy, H., Zhang, X., Domingues, C. M., Zhang, L., & Yuan, D. (2019a). Impacts of Basin-Scale Climate Modes on

Coastal Sea Level: a Review. *Surveys in geophysics*, *40*(6), 1493-1541. https://doi.org/10.1007/s10712-019-09562-8

Han, W., Stammer, D., Thompson, P., Ezer, T., Palanisamy, H., Zhang, X., Domingues, C. M., Zhang, L., & Yuan, D. (2019b). Impacts of basin-scale climate modes on coastal sea level: a review. *Surveys in geophysics*, *40*, 1493-1541.

Hauser, D., Tourain, C., Hermozo, L., Alraddawi, D., Aouf, L., Chapron, B., Dalphinet, A., Delaye, L., Dalila, M., Dormy, E., Gouillon, F., Gressani, V., Grouazel, A., Guitton, G., Husson, R., Mironov, A., Mouche, A., Ollivier, A., Oruba, L., . . . Tran, N. (2021). New Observations From the SWIM Radar On-Board CFOSAT: Instrument Validation and Ocean Wave Measurement Assessment. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(1), 5-26. https://doi.org/10.1109/TGRS.2020.2994372

Holgate, S. J., Matthews, A., Woodworth, P. L., Rickards, L. J., Tamisiea, M. E., Bradshaw, E., Foden, P. R., Gordon, K. M., Jevrejeva, S., & Pugh, J. (2013). New data systems and products at the permanent service for mean sea level. *Journal of Coastal Research*, *29*(3), 493-504.

Hughes, C.W., Fukumori, I., Griffies, S.M. *et al.*, 2019: Sea Level and the Role of Coastal Trapped Waves in Mediating the Influence of the Open Ocean on the Coast. *Surv Geophys* **40**, 1467–1492 .

Jean-Michel, L., Eric, G., Romain, B.-B., Gilles, G., Angélique, M., Marie, D., Clément, B., Mathieu, H., Olivier, L. G., & Charly, R. (2021). The Copernicus global 1/12 oceanic and sea ice GLORYS12 reanalysis. *Frontiers in Earth Science*, *9*, 698876.

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., & Balsamo, G. (2019). SEAS5: the new ECMWF seasonal forecast system. *Geoscientific Model Development*, *12*(3), 1087-1117.

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., Keeley, S. P. E., Mogensen, K., Zuo, H., & Monge-Sanz, B. M. (2019). SEAS5: the new ECMWF seasonal forecast system. *Geosci. Model Dev.*, *12*(3), 1087-1117. https://doi.org/10.5194/gmd-12-1087-2019

Kennedy, J. J., Rayner, N. A., Atkinson, C. P., & Killick, R. E. (2019). An Ensemble Data Set of Sea Surface Temperature Change From 1850: The Met Office Hadley Centre HadSST.4.0.0.0 Data Set. *Journal of Geophysical Research: Atmospheres*, *124*(14), 7719-7763. https://doi.org/https://doi.org/10.1029/2018JD029867

Kharin, V. V., & Zwiers, F. W. (2003). On the ROC score of probability forecasts. *Journal of Climate*, *16*(24), 4145-4150.

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., Van Den Dool, H., Saha, S., Mendez, M. P., & Becker, E. (2014). The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bulletin of the American Meteorological Society*, *95*(4), 585-601.

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., van den Dool, H., Saha, S., Mendez, M. P., Becker, E., Peng, P., Tripp, P., Huang, J., DeWitt, D. G., Tippett, M. K., Barnston, A. G., Li, S., Rosati, A., Schubert, S. D., . . . Wood, E. F. (2014). The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. *Bulletin of the American Meteorological Society*, *95*(4), 585-601. https://doi.org/https://doi.org/10.1175/BAMS-D-12-00050.1

Livezey, R. E., & Chen, W. (1983). Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev*, *111*(1), 46-59.

Long, X., Shin, S. I., & Newman, M. (2023). Statistical downscaling of seasonal forecasts of sea level anomalies for US coasts. *Geophysical Research Letters*, *50*(4), e2022GL100271.

Long, X., Widlansky, M. J., Schloesser, F., Thompson, P. R., Annamalai, H., Merrifield, M. A., & Yoon, H. (2020). Higher Sea Levels at Hawaii Caused by Strong El Niño and Weak Trade Winds. *Journal of Climate*, *33*(8), 3037-3059. https://doi.org/https://doi.org/10.1175/JCLI-D-19-0221.1

Long, X., Widlansky, M. J., Spillman, C. M., Kumar, A., Balmaseda, M., Thompson, P. R., Chikamoto, Y., Smith, G. A., Huang, B., Shin, C.-S., Merrifield, M. A., Sweet, W. V., Leuliette, E., Annamalai, H. S., Marra, J. J., & Mitchum, G. (2021). Seasonal Forecasting Skill of Sea-Level Anomalies in a Multi-Model Prediction Framework. *Journal of Geophysical Research: Oceans*, *126*(6), e2020JC017060. https://doi.org/https://doi.org/10.1029/2020JC017060

Lu, F., Harrison, M. J., Rosati, A., Delworth, T. L., Yang, X., Cooke, W. F., Jia, L., McHugh, C., Johnson, N. C., Bushuk, M., Zhang, Y., & Adcroft, A. (2020). GFDL's SPEAR Seasonal Prediction System: Initialization and Ocean Tendency Adjustment (OTA) for Coupled Model Predictions. *Journal of Advances in Modeling Earth Systems*, *12*(12), e2020MS002149. https://doi.org/https://doi.org/10.1029/2020MS002149

Mason, S. J., & Graham, N. E. (1999). Conditional probabilities, relative operating characteristics, and relative operating levels. *Weather and Forecasting*, *14*(5), 713-725.

May, C. L., Osler, M. S., Stockdon, H. F., Barnard, P. L., Callahan, J. A., Collini, R. C., Ferreira, C. M., Finzi Hart, J., Lentz, E. E., Mahoney, T. B., Sweet, W., Walker, D., & Weaver, C. P. (2023). Coastal effects. In A. R. Crimmins, C. W. Avery, D. R. Easterling, K. E. Kunkel, B. C. Stewart, & T. K. Maycock (Eds.), *Fifth National Climate Assessment*. U.S. Global Change Research Program. https://doi.org/10.7930/NCA5.2023.CH9

McIntosh, P. C., Church, J. A., Miles, E. R., Ridgway, K., & Spillman, C. M. (2015). Seasonal coastal sea level prediction using a dynamical model. *Geophysical Research Letters*, *42*(16), 6747-6753.

Miles, E. R., Spillman, C. M., Church, J. A., & McIntosh, P. C. (2014). Seasonal prediction of global sea level anomalies using an ocean–atmosphere dynamical model. *Climate dynamics*, *43*, 2131-2145.

Newman, M., & Sardeshmukh, P. D. (2017). Are we near the predictability limit of tropical Indo-Pacific sea surface temperatures? *Geophysical Research Letters*, *44*(16), 8520-8529.

Nicholls, R. J., Lincke, D., Hinkel, J., Brown, S., Vafeidis, A. T., Meyssignac, B., Hanson, S. E., Merkens, J.-L., & Fang, J. (2021). A global analysis of subsidence, relative sea-level change and coastal flood exposure. *Nature Climate Change*, *11*(4), 338-342. https://doi.org/10.1038/s41558-021-00993-z

Parker, B. B. (2007). Tidal analysis and prediction.

Penland, C., & Sardeshmukh, P. D. (1995). The Optimal Growth of Tropical Sea Surface Temperature Anomalies. *Journal of Climate*, *8*(8), 1999-2024. https://doi.org/https://doi.org/10.1175/1520-0442(1995)008<1999:TOGOTS>2.0.CO;2

Piecuch, C. G., Bittermann, K., Kemp, A. C., Ponte, R. M., Little, C. M., Engelhart, S. E., & Lentz, S. J. (2018). River-discharge effects on United States Atlantic and Gulf coast sea-level changes. *Proceedings of the National Academy of Sciences*, *115*(30), 7729-7734.

Pugh, D., & Woodworth, P. (2014). *Sea-level science: understanding tides, surges, tsunamis and mean sea-level changes*. Cambridge University Press.

Ray, R., Widlansky, M., Genz, A., & Thompson, P. (2023). Offsets in tide-gauge reference levels detected by satellite altimetry: ten case studies. *Journal of Geodesy*, *97*(12), 110.

Roberts, C., Calvert, D., Dunstone, N., Hermanson, L., Palmer, M., & Smith, D. (2016). On the drivers and predictability of seasonal-to-interannual variations in regional sea level. *Journal of Climate*, *29*(21), 7565-7585.

Ryan, H., & Noble, M. (2002). Sea level response to ENSO along the central California coast: how the 1997–1998 event compares with the historic record. *Progress in Oceanography*, *54*(1-4), 149-169.

Sardeshmukh, P. D., J. A. Wang, G. P. Compo, and C. Penland, 2023: Improving Atmospheric Models by Accounting for Chaotic Physics. *J. Climate*, **36**, 5569–5585

Shin, S.-I., & Newman, M. (2021). Seasonal Predictability of Global and North American Coastal Sea Surface Temperature and Height Anomalies. *Geophysical Research Letters*, *48*(10), e2020GL091886. https://doi.org/https://doi.org/10.1029/2020GL091886

Smith, D. M., Eade, R., & Pohlmann, H. (2013). A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction. *Climate dynamics*, *41*(11), 3325-3338. https://doi.org/10.1007/s00382-013-1683-2

Taherkhani, M., Vitousek, S., Barnard, P. L., Frazer, N., Anderson, T. R., & Fletcher, C. H. (2020). Sea-level rise exponentially increases coastal flood frequency. *Scientific reports*, *10*(1), 1-17.

Thompson, P. R., Widlansky, M. J., Hamlington, B. D., Merrifield, M. A., Marra, J. J., Mitchum, G. T., & Sweet, W. (2021). Rapid increases and extreme months in projections of United States high-tide flooding. *Nature Climate Change*, *11*(7), 584-590. https://doi.org/10.1038/s41558-021-01077-8

Tippett, M. K., & L'Heureux, M. L. (2020). Low-dimensional representations of Niño 3.4 evolution and the spring persistence barrier. *npj Climate and Atmospheric Science*, *3*(1), 24.

Toth, Z., Talagrand, O., & Zhu, Y. (2006). The attributes of forecast systems: A general framework for the evaluation and calibration of weather forecasts. *Predictability of Weather and Climate*, *584*(2006), 595.

van den Dool, H. (2006). *Empirical Methods in Short-Term Climate Prediction*. Oxford University Press. https://doi.org/10.1093/oso/9780199202782.001.0001

Vitousek, S., Barnard, P. L., Fletcher, C. H., Frazer, N., Erikson, L., & Storlazzi, C. D. (2017). Doubling of coastal flooding frequency within decades due to sea-level rise. *Scientific reports*, *7*(1), 1-9.

Wang, G., Ren, H.-L., Liu, J., & Long, X. (2023). Seasonal predictions of sea surface height in BCC-CSM1.1m and their modulation by tropical climate dominant modes. *Atmospheric Research*, *281*, 106466. https://doi.org/https://doi.org/10.1016/j.atmosres.2022.106466

Wang, J., Church, J. A., Zhang, X., & Chen, X. (2021). Reconciling global mean and regional sea level change in projections and observations. *Nature Communications*, *12*(1), 990. https://doi.org/10.1038/s41467-021-21265-6

Weisheimer, A., & Palmer, T. N. (2014). On the reliability of seasonal climate forecasts. *Journal of the Royal Society Interface*, *11*(96), 20131162.

Widlansky, M. J., Long, X., Balmaseda, M. A., Spillman, C. M., Smith, G., Zuo, H., Yin, Y., Alves, O., & Kumar, A. (2023). Quantifying the Benefits of Altimetry Assimilation in

Seasonal Forecasts of the Upper Ocean. *Journal of Geophysical Research: Oceans*, *128*(5), e2022JC019342. https://doi.org/https://doi.org/10.1029/2022JC019342

Widlansky, M. J., Marra, J. J., Chowdhury, M. R., Stephens, S. A., Miles, E. R., Fauchereau, N., Spillman, C. M., Smith, G., Beard, G., & Wells, J. (2017). Multimodel ensemble sea level forecasts for tropical Pacific islands. *Journal of Applied Meteorology and Climatology*, *56*(4), 849-862.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (Vol. 100). Academic press.

Wöppelmann, G., & Marcos, M. (2016). Vertical land motion as a key to understanding sea level change and variability. *Reviews of Geophysics*, *54*(1), 64-92. https://doi.org/https://doi.org/10.1002/2015RG000502

Wulff, C. O., Vitart, F., & Domeisen, D. I. V. (2022). Influence of trends on subseasonal temperature prediction skill. *Quarterly Journal of the Royal Meteorological Society*, *148*(744), 1280-1299. https://doi.org/https://doi.org/10.1002/qj.4259

Xie, D., Zou, Q.-P., Mignone, A., & MacRae, J. D. (2019). Coastal flooding from wave overtopping and sea level rise adaptation in the northeastern USA. *Coastal Engineering*, *150*, 39-58.

Xue, Y., Huang, B., Hu, Z.-Z., Kumar, A., Wen, C., Behringer, D., & Nadiga, S. (2011). An assessment of oceanic variability in the NCEP climate forecast system reanalysis. *Climate dynamics*, *37*(11), 2511-2539. https://doi.org/10.1007/s00382-010-0954-4

Zervas, C. E., Gill, S. K., & Sweet, W. W. V. (2013). Estimating vertical land motion from long-term tide gauge records.

Zuo, H., Balmaseda, M. A., Tietsche, S., Mogensen, K., & Mayer, M. (2019). The ECMWF operational ensemble reanalysis–analysis system for ocean and sea ice: a description of the system and assessment. *Ocean science*, *15*(3), 779-808.