Statistical-Dynamical Forecasting of Sub-Seasonal North Atlantic Tropical Cyclone Occurrence

Michael Maier-Gerber[*]

Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Karlsruhe, Germany

Andreas H. Fink

Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Karlsruhe, Germany

Michael Riemer

Institute for Atmospheric Physics, Johannes Gutenberg University, Mainz, Germany

Elmar Schoemer

Institute of Computer Science, Johannes Gutenberg University, Mainz, Germany

Christoph Fischer

Institute of Computer Science, Johannes Gutenberg University, Mainz, Germany

Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Karlsruhe, Germany

Benedikt Schulz

Early Online Release: This preliminary version has been accepted for publication in *Weather and Forecasting*, may be fully cited, and has been assigned DOI 10.1175/WAF-D-21-0020.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

Institute for Stochastics, Karlsruhe Institute of Technology, Karlsruhe, Germany

[*] *Corresponding author*: Michael Maier-Gerber, michael.maier-gerber@kit.edu

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

ABSTRACT

While previous research on sub-seasonal tropical cyclone (TC) occurrence has mostly focused on either the validation of numerical weather prediction (NWP) models, or the development of statistical models trained on past data, the present study combines both approaches to a statistical–dynamical model for probabilistic forecasts in the North Atlantic basin. Although state-of-the-art NWP models have been shown to lack predictive skill with respect to sub-seasonal weekly TC occurrence, they may predict the environmental conditions sufficiently well to generate predictors for a statistical model. Therefore, an extensive predictor set was generated, including predictor groups representing the climatological seasonal cycle (CSC), oceanic, and tropical conditions, tropical wave modes, as well as extratropical influences, respectively. The developed hybrid forecast model is systematically validated for the Gulf of Mexico and Central Main Development Region (MDR) for lead times up to five weeks. Moreover, its performance is compared against a statistical approach trained on past data, as well as against different climatological and NWP benchmarks. For sub-seasonal lead times, the CSC models are found to outperform the NWP models, which quickly loose skill within the first two forecast weeks, even in case of recalibration. The statistical models trained on past data increase skill over the CSC models, whereas even greater improvements in skill are gained by the hybrid approach out to week five. The vast majority of the additional sub-seasonal skill in the hybrid model, relative to the CSC model, could be attributed to the tropical (oceanic) conditions in the Gulf of Mexico (Central MDR).

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0020.1.

1. Introduction

For decades, there has been a parallel development of predictions for individual tropical cyclones (TCs) made by operational forecast centers for lead times of a few days on the one hand, and seasonal predictions of integrated TC activity on the other. This coexistence is due to the sub-seasonal predictability gap (Vitart et al. 2012; Robertson et al. 2020), which has raised broad attention and efforts to bridge only in recent years. Because of the lack of skillful models, potential sources for sub-seasonal predictability of tropical cyclone activity have become an increased research focus with the growing understanding of various modes of intraseasonal to interannual variability (Camargo et al. 2019), such as the Madden-Julian Oscillation (MJO) and the El Niño Southern Oscillation (ENSO). Since nowadays numerical weather prediction (NWP) models are often integrated to sub-seasonal or seasonal forecast horizons, these atmospheric modes of variability have been shown to impact sub-seasonal NWP forecasts for TC activity in many oceans (e.g., Vitart 2009; Belanger et al. 2010; Camp et al. 2018). Several studies have systematically evaluated these models in terms of predictive skill for different TC occurrence measures (Lee et al. 2018, 2020; Gregory et al. 2019). Lee et al. (2018) found that the Sub-seasonal to Seasonal (S2S; Vitart et al. 2017) models generally have little to zero skill in predicting TC occurrence from week two on for all basins relative to climatological forecasts. For the North Atlantic, they stated that actual and potential model skills are very close, suggesting that hardly any improvement can be achieved with current NWP models.

Inspired by the example of numerous statistical forecast models for seasonal forecasting, Leroy and Wheeler (2008) followed a different approach and developed logistic regression models based on past data to produce probabilistic forecasts of weekly TC genesis and occurrence in four zones of the Southern Hemisphere up to seven weeks in advance. Comparing against ECMWF model predictions, Vitart et al. (2010) identified the statistical approach from Leroy and Wheeler (2008) to perform better from week two on. They also compared against a simple recalibrated version of the ECMWF forecasts, as well as against an average of the statistical and the recalibrated ECMWF model predictions, which further improved skill. For the North Atlantic, Slade and Maloney (2013) used the successful approach from Leroy and Wheeler (2008) as a blueprint and generated basin-wide forecasts on the basis of a predictor set adopted to that ocean basin. Although using the same statistical approach, the predictor sets used in Leroy and Wheeler (2008), Vitart et al. (2010), and Slade and Maloney (2013) slightly differ. While all have in common that they provide a climatological seasonal cycle, and the two Real-time Multivariate MJO (RMM; Wheeler and Hendon 2004) indices for model training, they vary in which and how the oceanic modes of variability are represented.

Even though Lee et al. (2018) concluded that the S2S models lack skill to forecast North Atlantic sub-seasonal TC genesis, these models may be able to predict sub-seasonal environmental conditions favourable for TC genesis to a sufficient degree, so that predictors can be generated and fed into statistical models. Such a statistical–dynamical (or hybrid) forecast model is thought to combine the strengths of each individual model, and thus to increase model skill. This has been demonstrated for seasonal predictions of basin-wide TC activity in several oceans (Klotzbach et al. 2019, 2020). For sub-seasonal leadtimes, Qian et al. (2020) developed a hybrid model for basin-wide TC genesis counts, to then derive TC track distributions for the Western North Pacific. The present study aims to develop a hybrid model for the North Atlantic, using a variety of climatological, oceanic, tropical, and extratropical predictors known to precondition and modulate environments that are prone to TC occurrence. In addition to the predictors used in Leroy and Wheeler (2008), Vitart et al. (2010), and Slade and Maloney (2013), further potentially relevant predictors, related to the genesis potential index (GPI; Emanuel and Nolan 2004), tropical waves, and extratropical PV dynamics, will be included.

While different measures are defined and used to describe TC activity in the literature, we here focus on the binary aspect of weekly TC occurrence. Instead of basin-wide forecasts, we employ a gridded framework to gain insight into subregional differences. Beyond the development of the hybrid approach, an important contribution of this paper is to compare a hierarchy of different model types in terms of predictive skill in a systematic way. To the authors knowledge, the present study is the first to compare subregional North Atlantic TC occurrence forecasts out to week five of i) a statistical–dynamical approach with ii) a purely statistical approach (as in Slade and Maloney 2013), iii) different climatological models, as well as iv) (un)calibrated S2S models at once. In the following, the datasets and methods are described in section 2. Section 3 then introduces the benchmark models, before the individual predictors for the two statistical approaches are motivated and developed in section 4. The strategy of how optimal predictor sets are determined, and how statistical models are formulated, trained, and validated is summarized in section 5. Results of the systematic model comparison are finally presented in section 6, followed by a summary and conclusive remarks in section 7.

2. Data and methods

a. Target variable: TC occurrence

The basis for deriving the TC occurrence is the International Best Track Archive for Climate Stewardship (IBTrACS; Knapp et al. 2010, 2018) dataset version 4. To account for TC occurrence, cyclones are required to be tropical in nature and to exceed at least tropical storm strength ($\geq$ 34 kn). Although the IBTrACS dataset comes with a 3-hourly temporal resolution, only 0000 UTC instances of cyclone track positions are taken into account to allow for a systematic comparison with the lowest temporally resolved benchmark model, the S2S TC tracks (cf. section 3b). Figure 1a shows the North Atlantic cyclone positions, that fulfill the stated criteria for the periods used for model validation, training of the statistical models, and for generating the climatological models, respectively.

To take account of the reduced predictability on sub-seasonal timescales, TC occurrence (hereafter alternatively referred to as 'target variable') is created by means of a coarser spatio-temporal evaluation, which is over periods of one week and within a certain spatial area. For a given forecast week, a grid point is considered to feature TC occurrence, if at least one TC occurs within a radial distance of 7.5°. Different radii have been tested, but 7.5° has been chosen as a compromise between skill and usefulness of the prediction. Using a 1.5° × 1.5° grid, the circular evaluation domains partially overlap. Based on the dichotomous target variable, Figure 1b presents a map of the resulting relative frequencies of TC occurrence, which can also be interpreted as an occurrence density plot.

b. Predictor variables for the statistical models

The difference between the statistical-dynamical approach and the purely statistical approach developed is merely in the underlying data, from which predictors are generated. The purely statistical models are trained on ERA5 (Hersbach et al. 2020) data, whereas predictors for the statistical-dynamical models are generated from S2S ECMWF ensemble reforecasts. For the latter, we use model version dates from the 2018 North Atlantic Hurricane season (Jun.-Nov.), which means that the corresponding reforecasts belong to the 1998–2017 period. While the ERA5 dataset was produced by model version Cy41r2 of the Integrated Forecasting System (IFS) – ECMWF's atmospheric model and data assimilation system –, the S2S reforecasts were based on Cy45r1.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

Despite some changes (e.g., data assimilation, atmosphere–ocean coupling, or parameterization schemes), the horizontal and vertical resolution of the IFS model remained unchanged between the two cycles, and the fact, that both datasets are based on the IFS model, allows to more clearly attribute differences in skill to differences in model approaches.

The S2S reforecasts are produced twice per week (Mondays and Thursdays) with one control plus 10 perturbed forecasts, ranging out to 46 days. Originally calculated with a horizontal grid spacing of 16 km for the first 15 days and 31 km afterwards, S2S model output is archived with daily values at 0000 UTC on a regular $1.5° \times 1.5°$ grid, which is considerably coarser compared to ERA5. For the sake of consistency, both datasets are therefore used with this coarser grid spacing and temporal resolution. Since only basic fields are available from the S2S dataset, potential vorticity (PV) was calculated from the available pressure levels 50, 100, 200, 300, 500, 700, 850, 925, and 1000 hPa, using the approximation from Bluestein (1993). To ensure that the S2S-based predictors are not subject to biases, a mean bias correction was applied to all variables, from which predictors were directly generated. Biases were calculated with respect to the 1979–2018 ERA5 period as a function of day of year, forecast time, and location. Since the basic assumption for a forecast ensemble is the independence and interchangeability of the individual members, biases are not regarded to be a function of the ensemble member. Undesirable fluctuations in the seasonal cycle of the biases are smoothed out by applying a 31-day moving average.

c. Tropical wave filtering

Since tropical waves are characterized by their propagating nature in space and time, there is no unique approach to identify and analyze those in a given dataset, although plenty methods have been proposed, each having its pros and cons. The current study applies the wave filtering method described in Janiga et al. (2018), who applied the zero-padding strategy from Wheeler and

Weickmann (2001) to the end of S2S reforecasts, which are prepended by reanalysis data. The time series to be filtered have a total length of four years – two years of reanalysis data, and two years for S2S reforecast data plus zeros (Fig. 2a). Wave-related predictors derived for the purely statistical approach are derived from the same time series but with the S2S reforecasts replaced by zeros (Fig. 2b). To more specifically evaluate sub-seasonal signals during filtering, the first four harmonics of the 1979–2018 annual cycle are calculated from ERA5 and subtracted from all non-zero portions of the composed time series. As illustrated in Figure 2, the first and last years of the four-year time series are tapered to zero using a split-cosine-bell to mitigate that filtering results suffer from spectral leakage.

The filtering method is applied to horizontal wind divergence at 200 and 850 hPa for every latitude and ensemble member separately, to filter for five wave types (listed in Table 1) in frequency–wavenumber domain. The filter windows used are identical to the ones proposed in Janiga et al. (2018) and are larger than those applied in many climatological studies (e.g.; Wheeler and Kiladis 1999) to take into account that wave propagation characteristics predicted by the S2S model may differ from their real-world counterparts, in particular on sub-seasonal timescales. Even though the filter windows were developed for the solutions of the shallow water equations, the approach is here applied for latitudes ranging from the equator into the extratropics for pragmatic reasons.

d. Forecast verification metrics

A model is said to be calibrated (or reliable) if the forecast probabilities match the observed relative frequencies, for example, if we issue a forecast probability of 20% 100 times, the event should occur about 20 times. For a set of *N* forecast-observation pairs, calibration can be assessed using the mean Brier score (BS; Brier 1950), which is defined as

$$BS = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{p}_i)^2, \tag{1}$$

where y_i is the observation (either 0 or 1), and $\hat{p}_i$ the forecast probability (between 0 and 1) of the i-th instance. The BS ranges from 0 to 1, is negatively oriented (i.e., lower is better), and a strictly proper scoring rule, meaning that the expected score is minimized by the true underlying distribution of the observation (Gneiting and Raftery 2007). To assess skill gain relative to a reference model, here chosen as the mean seasonal climatology (MSC, see section 3a), the Brier skill score (BSS) is calculated, given by

$$BSS = 1 - \frac{BS}{BS_{MSC}}. \tag{2}$$

Thus, a model to be validated has better, no, and less skill compared to the MSC model if the BSS becomes greater, equal, and less than zero, respectively.

3. Benchmark models

An integral part of model development is to compare a newly generated model with those that are well-established and/or different in their approach. To justify the application of a new model, it should perform better than the models chosen to serve as benchmark. With climatological and NWP models, two distinct types of benchmark models are employed in the following to put into relation the performance of the statistical models developed.

a. Climatological forecasts of TC occurrence

Climatological models are used as the first type of benchmark to allow for a comparison with predictions based on long-term statistics of TC occurrence, i.e. on its climatology. Because those statistics are calculated over a set of past realizations drawn from the underlying distribution of the target variable, climatological forecasts are inherently independent of the current state of the atmosphere. Moreover, they are unbiased if trends and/or regime changes are negligible. If so, there are no restrictions regarding lead time, and forecasts are thus independent of forecast week. The climatological models used here are derived from the IBTrACS dataset for the period 1968–2017 (see Fig. 1a, orange and blue dots). Because a complete and consistent TC monitoring was only possible since the beginning of the satellite era, seasons earlier than 1968 are not considered for calculating the climatologies. The simplest approach to generate a climatological statistic is to average TC occurrence over the 50 North Atlantic Hurricane seasons considered. This approach yields a mean seasonal climatology (MSC), where constant forecasts are predicted throughout the season. A more adaptive strategy to take into account seasonal variations is to average over years for every day of year separately, resulting in a climatological seasonal cycle (CSC).

Seasonal fluctuations evident from the CSC example in Fig. 3 indicate that the 50-year period is not sufficient to generate a robust climatology, since one would expect the observed relative frequency to not vary much for neighboring days in the year. To mitigate the adverse effect of too small sample sizes, a smoother and more representative CSC (hereafter referred to as *CSCopt*) was constructed by applying a moving average. The optimal window length at every grid point was identified by maximizing the Pearson correlation with the target variable. Various commonly used weighting kernels were tested, but a simple uniform weighting turned out to yield the highest correlations overall.

b. S2S forecasts of TC occurrence

To compare with predictions directly obtained from a state-of-the-art NWP model, a second type of benchmark is created by calculating probabilities for TC occurrence from all 0000 UTC instances of the TC tracks identified in the S2S ECMWF 1998-2017 ensemble reforecasts (hereafter referred to as *S2STC*). These tracks are publicly available for download[1] and are based on the TC detection algorithm described in Vitart and Stockdale (2001). They also present details on how precision of TC locations, identified in the relatively coarse S2S model output, is further increased. The 0000 UTC TC locations are finally composed to tracks by applying the tracking technique presented in Van der Grijn et al. (2005). Since TCs are analyzed in the S2S reforecasts, temporal resolution, number of ensemble members, and forecast range are consistent with the original model output (cf. section 2b). Because the coarse resolution of the S2S model output results in a low TC intensity bias, the lowered threshold of 24 kn, suggested by Lee et al. (2018), is also used here to define tropical storm strength.

Since S2STC forecasts are frequently not calibrated, we have tested different techniques to correct for potential miscalibration. A non-parametric method, called isotonic distributional regression (IDR; Henzi et al. 2019), turned out to perform best for that purpose. Based on the natural assumption that a higher forecast probability is associated with a higher event frequency, IDR learns a step-function that is used to transform the S2STC forecasts to calibrated probability forecasts[2]. To increase robustness, forecasts from all grid points of a given validation subregion are pooled for training the isotonic regression, which is then applied to every grid point separately. For a quality assessment, the calibrated forecasts (hereafter referred to as *S2STCcal*) will be assessed in section 6a.

4. Predictor development and analysis

Because training of statistical models requires a set of relevant predictors, this section presents an expert selection of predictors from different categories. To contrast the two statistical approaches,

[1] ftp://s2sidx:s2sidx@acquisition.ecmwf.int/TCYC

[2] In the application on probability forecasts, IDR is equivalent to isotonic regression, a common approach for calibration of probabilities in the machine learning literature (e.g. in Guo et al. 2017).

predictor generation is motivated and described for the statistical–dynamical approach, followed by a description of how an equivalent set is constructed for the purely statistical approach. The selection of predictors is neither meant to be complete nor the most sophisticated way of how predictors can be generated, but constitutes a solid foundation for statistical model development.

a. Climatological predictor derived from the IBTrACS dataset

Given the chaotic nature of the atmosphere, the skill of any model can be expected to steadily converge to the skill of the corresponding climatological model for long enough lead times. For this reason, the CSCopt model (see section 3a) is not only deployed as a benchmark model, but also constitutes the base predictor for the statistical models. This ensures that they are able to at least exploit information from intraseasonal variations as sort of a "fail-save", in case they cannot gain any skill from the data of the NWP-based predictors during model training, due to insufficient signal-to-noise ratios on sub-seasonal lead times (Scheuerer et al. 2020). Any positive differences in skill relative to the CSCopt can thus be attributed to the added value of the NWP-based predictors.

Figure 4 presents the Pearson correlation coefficient ρ between the CSCopt predictor and the target variable, calculated from all forecast–observation pairs of the 1998-2017 seasons, separately for every grid point. Since the predictor and the target variable have the underlying dataset in common, the season to be forecast is left out, when generating the CSCopt predictor. Correlations are found to be positive throughout the entire basin, and significant for almost all grid points, where forecast models are developed (cf. Fig. 1, red contour). Peak correlation values of up to 0.5 are located slightly north of the center of the so-called Main Development Region (MDR; 80°W–20°W, 10°N–20°N), slowly decaying towards the US east coast. Because this predictor is independent of the forecast week, the described correlation patterns are valid for all forecast weeks considered. Given these correlations, the CSCopt predictor is a good starting point for the predictor set.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

b. S2S reforecast predictors for the statistical-dynamical approach

Since S2S ECMWF reforecasts are run in ensemble mode, we want to make use of the valuable information on forecast uncertainty. For each S2S-based predictor variable, we therefore calculate the mean and standard deviation to represent the first and second statistical moments of the ensemble's distribution. Providing those as separate predictors, the statistical models should learn primarily from the predictive signals associated with the ensemble means in case where ensemble uncertainties remain sufficiently low. However, the standard deviation predictors become increasingly important when they exceed the standard deviation of the ensemble means of all training instances, since the information from the ensemble mean predictors becomes less relevant in such cases.

Predictors are constructed from the fields forecasted by the S2S model in two ways, based on whether they represent an immediate (local predictor) or potentially lagged (remote predictor) influence. For a local predictor, at every grid point, the corresponding S2S forecast field is averaged over the same week used for the target variable, and within a radius of 7.5°, to be consistent with the integral perspective on weekly TC occurrence motivated in section 2a. In contrast, when constructing remote predictors, using the same forecast week for the S2S forecast fields as for the target variable does not necessarily yield the optimal link. For instance, a week-three TC occurrence forecast could also be more strongly correlated with predictors constructed from S2S fields of week one or two, respectively. The optimal link is essentially a trade-off between reduced S2S forecast errors (when the chosen predictor week is closer to the initialization of the S2S forecast), and smaller time lags (when the chosen predictor week is closer to the target forecast week), to more directly link the physical relationship. For each remote predictor of the ensemble mean, we determined the optimal S2S forecast week by maximizing the Pearson correlation with the target forecast week. To more easily discuss the mapping in the following presentation of the constructed remote predictors, the correlations calculated at every gridpoint were averaged across the basin, before being applied to every grid point again. The results for each remote predictor of the ensemble mean were likewise applied to the corresponding predictor of ensemble standard deviation.

1) Oceanic predictors

The local SSTs play a crucial role for TC genesis (Palmen 1948) by providing the energy resource for the intensification and maintenance of the convectively driven secondary circulation (Gray 1968) through wind-induced surface heat exchange (WISHE; Emanuel 1986). Since SST data is not available over land, predictors for mean and standard deviation of local SST are only calculated and considered at grid points, where at least one SST value is given within the 7.5° radius. Figure 5a shows positive Pearson correlations for the mean predictor covering most of the North Atlantic basin, with a maximum at the northern edge of the central MDR.

Because most North Atlantic TCs form in the MDR, SSTs in this region are known to modulate basin-wide interannual TC activity (Shapiro 1982; Goldenberg and Shapiro 1996). Besides the immediate importance of local SST predictors, we therefore generate and include mean and standard deviation predictors of the SSTs averaged in the MDR. Since these predictors represent remote influences at all grid points outside the MDR, a pre-analysis was conducted to determine the optimal S2S predictor forecast week for each target forecast week. It turned out that spatio-temporal immediacy weighs more heavily than possible S2S model errors, since correlations are highest when the forecast week of SST and the target forecast week are the same. The identified correlation pattern reflects that higher MDR SSTs lead to an increased probability in TC occurrence for the vast majority of the grid points (Fig. 5b). The area with the highest correlations of greater

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

than 0.3 is found just northwest of the central MDR, which likely reflects the intensifying effect of high MDR SSTs on TC precursors originating over or close to West Africa.

Beyond basin-internal predictors, remote effects of SST via teleconnections are also well known. In contrast to MDR SSTs, eastern equatorial Pacific SSTs associated with ENSO are typically anticorrelated with TC activity during El Niño phases, and vice versa during La Niña (Goldenberg and Shapiro 1996). We analyzed SST predictors for the commonly defined Niño 1+2, Niño 3, and Niño 3.4 regions, but since the predictor–target Pearson correlations for Niño 1+2 were discernibly higher than for Niño 3 and Niño 3.4, respectively, Niño 1+2 was used to represent the ENSO state in this study. Although a time lag for the choice of the optimal predictor forecast weeks is expectable due to the remote influence, a pre-analysis for this region revealed that predictor and target are most strongly correlated when forecast weeks are the same. Apart from generally weaker correlations and the opposite sign, the close resemblance of the correlation patterns for the mean Niño 1+2 predictor (Fig. 5c) and mean MDR SST (Fig. 5b) underpins the ENSO teleconnection effect on MDR environmental conditions, identified by Gray (1984), and thus the potential value for including this remote predictor type.

2) Tropical predictors

In addition to the oceanic predictors, TC occurrence responds to a variety of atmospheric factors, which are known to be necessary for preconditioning the environment, in which a TC is likely to form and self-organize. Because the genesis potential index (GPI; Emanuel and Nolan 2004) was designed to assess near-storm environmental conditions, we created local predictors based on the terms contributing to the GPI, viz. 850-hPa absolute vorticity, 700-hPa relative humidity, 200–850-hPa vertical shear, and potential intensity. The latter, however, could not be calculated from the S2S database.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

Because a TC is characterized by a local absolute vorticity maximum, a zonal band of significant positive Pearson correlations for the week-four mean absolute vorticity predictor spans from the West African coast to the Gulf of Mexico, along the classical track of TCs initiated by African Easterly waves (Fig. 6a). This band is connected with an extension into the northeast Atlantic. Even though the correlation structure for the mean relative humidity predictor is similar to the one for absolute vorticity, variability within the zonal band is larger with a smaller maximum in the western Gulf of Mexico, and a pronounced maximum just west of the African coast (Fig. 6b). The latter is likely to be partly associated with the dryness of the Saharan Air Layer (SAL), which was found to impede TC genesis and intensification primarily over the eastern North Atlantic by facilitating convection-suppressing downdrafts (Dunion and Velden 2004). As expected and unlike the previous two GPI components, the detrimental effect of vertical wind shear results in anti-correlation, with highest absolute values in the MDR and the western Gulf of Mexico (Fig. 6c).

Furthermore, tropical waves have been shown to impact TC genesis by modulating their environmental conditions (e.g.; Frank and Roundy 2006). Frank and Roundy (2006) identified significant contributions of tropical waves up to one month prior to TC genesis, and highlighted the potential of these waves for statistical modeling. Therefore, we want to exploit this potential and filtered for tropical wave modes as described in section 2c. Although tropical waves are tied to the equator, their influence can be attributed to TC formation at latitudes beyond 30°N (Schreck III et al. 2012, see their Fig. 7). Given this remote link in the context of the fact that tropical waves and TCs are typically non-stationary, which makes it difficult to design predictors, we follow a pragmatic approach by generating local predictors from the latitude-wise filtered 200-hPa divergence squared. The squaring yields an activity measure, which proved to be more skillful compared to providing the phase information (non-squared). The 200-hPa level was preferred over 850 hPa due to higher correlation coefficients. The resulting predictor–target correlation coefficients are predominantly negative for most wave types (Fig. 7a-d), meaning that a reduced upper-level wave activity facilitates TC occurrence. The positive correlations associated with Mixed Rossby-gravity/tropical depression (MRG/TD) waves (Fig. 7e) can be partly explained by the fact that TCs project onto the filtering window used to define this wave type (Schreck III et al. 2011, see their Fig. 6).

The MJO is known to modulate North Atlantic TC activity (Maloney and Hartmann 2000), and thus has been used in purely statistical models for sub-seasonal TC occurrence before (e.g., Leroy and Wheeler 2008; Slade and Maloney 2013). As an alternative to the MJO-filtered local predictors, S2S ECMWF ensemble reforecasts of the more commonly used RMM indices were downloaded[3] to define MJO remote predictors. RMM indices are often used to distinguish between eight circumglobal phases of MJO-related convective activity, of which phase 2 (6+7) leads to significantly enhanced (reduced) North Atlantic TC activity (Klotzbach 2014; Camargo et al. 2009). However, an additional inclusion of the RMM predictors in the statistical-dynamical approach for testing purposes did not yield any further notable skill increase. Therefore, the MJO-filtered local predictors were used for statistical model development in the following, but not the RMM indices. Note that the lack of additional improvements does not contradict the modulation of TC activity by the RMM indices, which was previously documented in the ECMWF S2S model (Vitart 2009; Lee et al. 2018, 2020). Rather, it likely indicates that the predictive skill is covered by the local predictors already, which are modulated by the MJO through teleconnections.

3) Extratropical predictors

In recent years, a link between extratropical Rossby wave breaking (RWB) and North Atlantic TC activity has been revealed and accounted for another source to alter vertical shear and moisture,

[3] ftp://s2sidx:s2sidx@acquisition.ecmwf.int/RMMS

especially in the MDR (Zhang et al. 2016, 2017; Wang et al. 2020). RWB events typically yield a PV streamer, which often penetrates into the (sub)tropical regions. Papin et al. (2020) calculated a climatology for North Atlantic PV streamers, and found that a measure for climatologically standardized PV anomalies, integrated over frequency and area of the identified PV streamers, correlates better to TC activity than the individual measures alone. Even though not considering individual PV streamer objects, but gridpoint-wise averages within a 7.5° radius, we build on this finding of a stronger link when using an integral perspective, and generate local mean and standard deviation predictors for 200–500-hPa layer-averaged PV. These predictors are meant as a proxy for the integral effect of the presence of upper-level PV features that can influence TC occurrence in two ways. Due to their narrow shape, PV streamers typically feature high PV gradients, and hence high vertical shear, posing a detrimental environment for TC occurrence. On the other side, the PV streamer can spawn a low-level baroclinic precursor disturbance, which can undergo tropical transition once the upper-level PV gets diabatically redistributed (Davis and Bosart 2003; Maier-Gerber et al. 2019). However, because in this scenario TC occurrence takes place only after the vertical wind shear associated with the PV streamer is reduced to a sufficient degree (Davis and Bosart 2004), the adverse character of high PV is prevailing over the preceding supporting effect. This is confirmed by the negative correlations in Fig. 8, which are strongest in the northeastern edge of the MDR and along the US east coast, consistent with the stronger negative correlations found for the western basin by Zhang et al. (2017).

c. ERA5 predictors for the purely statistical approach

As opposed to the S2S-based predictors used for the statistical-dynamical approach, analogous ERA5-based predictors are generated for the purely statistical approach. This means that the S2S ensemble mean and standard deviation predictors are replaced by single predictors derived from

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

ERA5 data. Since NWP forecasts are not considered in this approach, the mean bias corrections as well as the pre-analyses for determining optimal predictor weeks are no longer required. Instead, predictor fields are averaged over the week before the date the S2S reforecast was initialized.

5. Statistical model development

a. Logistic regression

If the target variable is dichotomous, being either one or zero (i.e., TC occurrence or non-occurrence), logistic regression models (Hastie et al. 2009) are commonly trained to map linear combinations of continuous predictor variables to a probability via the so-called logit function. Given the training data $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, N$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iP})^T$ contains the values of the P predictors and y_i is the corresponding observation for a given instance i, the logistic regression model is formulated as

$$\hat{p}_i(\beta_0, \boldsymbol{\beta}) = logit^{-1}\left(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}\right) = \frac{1}{1 + \exp\left(-\beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta}\right)}, \tag{3}$$

where $\hat{p}_i$ is the estimated probability of the target variable instance y_i being one, β_0 is the intercept, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)^T$ the vector including the regression coefficients of the predictors. Using the LIBLINEAR solver (Fan et al. 2008), we estimate the coefficients based on the following problem:

$$\min_{\beta_0, \boldsymbol{\beta}} \quad \frac{1}{2}\boldsymbol{\beta}^T \boldsymbol{\beta} - \sum_{i=1}^{N} \left(y_i \log\left(\hat{p}_i\right) + (1 - y_i) \log\left(1 - \hat{p}_i\right)\right). \tag{4}$$

The second term corresponds to maximum likelihood estimation, the first to a l_2-penalty, which keeps the coefficients of the predictors small and thus prevents the model from overfitting. The minimization is stopped, if either the difference between the losses of two consecutive iterations drops below a tolerance of 10^{-4} , or a maximum number of 100 iterations is reached. To support faster convergence of solutions for model coefficients, predictors are standardized on the respective training set.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

b. Validation strategy and training of the statistical models

A systematic comparison of the different model approaches requires a common strategy for validation. While forecasts from the climatological benchmark models can be issued every day, the NWP-based benchmark model and the predictors for the logistic regression models rely on the twice-weekly run and disseminated S2S ECMWF forecasts, thus, posing the stronger limitation to a validation dataset. Starting from each of these S2S reforecast initialization dates, for every model, forecasts are generated for the first five consecutive weeks, i.e. day 0–6, 7–13, ..., 28–34. However, forecasts are only considered for validation if the middle of the respective forecast week falls into the North Atlantic Hurricane season, which runs from 1 June to 30 November. This yields a total of 1040 (52 reforecasts per season $\times$ 20 seasons) validation instances, for which S2S ECMWF reforecasts are available. In contrast to the NWP-based benchmark models, the climatological and logistic regression models require a training dataset that is independent of the validation dataset. To fully exploit the relatively small number of S2S reforecasts for both purposes, a 20-fold cross-validation (CV) is applied, so that every season can be successively validated, while the statistical models are being trained on the remaining 19 seasons. To avoid training statistical models with too great imbalances between TC occurrence and non-occurrence in the target variable, a fraction of at least 1 % is required to feature TC occurrence. This is the case for the Gulf of Mexico and Central MDR subregions as can be seen from the red contour enclosing the orange boxes in Fig. 1b. Because the climatological models (and thus the base predictor) share the underlying dataset with the target variable, the CV strategy necessitates the climatologies to be calculated separately for every fold, leaving out the data of the season to be forecast and validated. Although a separate statistical model is trained for every gridpoint and target forecast week, the generated forecasts are

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0020.1.

pooled for each of the two subregions, to allow for more solid conclusions during the validation discussed in section 6.

c. Sequential predictor selection

Training a logistic regression model on the full variety of predictors developed and motivated in section 4 does not necessarily lead to the best predictive performance. Optimal predictor subsets for the statistical-dynamical and purely statistical approach, respectively, are therefore determined using a sequential forward predictor selection. This selection process is conducted separately for the Gulf of Mexico and Central MDR subregions, and gridpoints are pooled for each subregion to make selections more robust. An overview of the potential predictor pools, from which the two approaches can choose, is presented in Fig. 9. To guarantee that the logistic regression models do not perform worse than the climatological benchmark models, the CSCopt predictor is kept fixed, a priori. This initial minimal subset is then extended by the one predictor that minimizes the average Akaike information criterion (AIC; Hastie et al. 2009; Akaike 1974) of a 5-fold CV on the training period. For a logistic regression model with P predictors, the AIC is defined as

$$AIC = -2\frac{LL}{N} + 2\frac{P}{N}, \tag{5}$$

where *LL* is the binomial log-likelihood based on *N* forecasts and the corresponding observations. We chose AIC as our scoring metric since it reduces overfitting by penalizing larger numbers of predictors, in addition to the term for the model's performance. The extension of the subset is repeated until all candidate predictors are integrated. Then, the optimal subset of predictors is finally identified by the lowest AIC achieved. This forward selection is preferred over a backward selection (i.e., successively removing predictors) to keep the number of optimal predictors as small as possible but as large as necessary. Similar to Leroy and Wheeler (2008) and Slade and Maloney

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

(2013), we first had performed a BS-based predictor selection in a pre-analysis on the full dataset. But since predictors should not be selected based on the data the models are validated on, the selection scheme was instead integrated in the 20-fold CV, such that predictors are chosen on the training data alone. Due to this change, the skill of the statistical models drastically decreased, but could be restored by pooling the gridpoints for each subregion and identifying optimal subsets based on the AIC. Hence, 20 predictor subsets are obtained that are found to be highly consistent, being in complete agreement for the central MDR, and differing in only one predictor at week two for the Gulf of Mexico. Because the focus of this study is on the development of a hybrid model, and the comparison of the distinct model approaches, the reader is referred to the supplement material for a brief overview and discussion of the predictor selection results.

6. Model comparison

a. Reforecast reliability

Before all models are validated in terms of skill, the joint distributions of the dichotomous target variable and the predicted probabilities should be analyzed to identify and, if necessary, correct for forecast biases. A common visual tool to easily identify conditional and unconditional biases are reliability diagrams (Sanders 1963; Wilks 2011), which display such joint distributions factorized into model reliability (calibration curve) and refinement (histogram). We here use the recently developed CORP approach (Dimitriadis et al. 2021), which has the advantage, among others, of providing optimally binned and readily reproducible diagrams. Figure 10 shows CORP reliability diagrams for the Gulf of Mexico and Central MDR week-four forecasts to represent the sub-seasonal time scale. Biases, however, are qualitatively similar for the other forecast weeks. For both subregions and all models, forecast probabilities tend to be generally very low, consistent with the extreme nature of TCs, leading to low relative frequencies of TC occurrence in the target variable (cf. Fig. 1b). Thus, the model predictions can be made only with low confidence as the forecast probabilities are distributed mainly around the mean relative frequency of the target variable. Since the rareness of TC occurrence is given by nature, the only remedy would be to increase the evaluation radius beyond 7.5°, which, however, would inevitably lead to an also undesirable larger uncertainty in spatial interpretation. However, it can be stated that the logistic regression models can predict with slightly higher confidence compared to the benchmark models.

A model is well calibrated (or reliable) when the forecasted probabilities match the observed relative frequencies. Miscalibration can thus be visually assessed through deviations of the coloured curves from the diagonal. The first thing to notice is that all models are more reliable for low forecast probabilities than for higher ones, which is consistent with the refinement distributions discussed before. The underforecasting situation (TC non-occurrence bias) of the CSCopt model is likely to result from a reduced TC occurrence in the 1968-1997 period, which was used to extend the 1998-2017 validation period for calculating more robust climatologies. However, since the CSCopt is also a base predictor for the logistic regression models, it has no competitive disadvantage when evaluating model skill. The S2STC model similarly underforecasts the low forecast probabilities, but overforecasts the few high forecast probabilities, which results in a general overconfidence. To correct for this conditional bias, this particular NWP-based model is calibrated using the IDR method described in section 3b. The S2STCcal follows the diagonal quite well for low forecast probabilities, and thus generates much more reliable forecasts. Since logistic regression is known to yield well-calibrated forecasts, the calibration curves for the two approaches of logistic regression models are well-aligned with the diagonal for low forecast probabilities. The increasing deviations with higher forecast probabilities are likely due to the few samples, which are obviously insufficient for generalization. Overall, sub-seasonal forecasts of the logistic regression models, with a slightly better calibrated statistical-dynamical approach for higher forecast probabilities, are more reliable than the benchmark forecasts.

b. Reforecast model skill

1) Comparison of different model approaches

Figure 11 shows a comparison of the BSS as a function of forecast week for the different model types validated in the Gulf of Mexico and Central MDR subregions. Since climatological forecasts are independent of the forecast week, the BSS also does not change with lead time. Considering that the MSC is used as reference, the positive BSS for the CSC and CSCopt models indicate that the ability to simulate seasonal variations is rewarded. The improvement in skill, however, exhibits remarkable subregional differences, as can be seen by CSC BSSs three times stronger (about 15 % vs. less than 5 %) for the Central MDR compared to the Gulf of Mexico. This is due to the fact that TC occurrence in the MDR is often associated with African Easterly Waves, which are subject to a more distinct seasonal cycle (Thorncroft and Hodges 2001). The CSC BSSs are further enhanced when correcting for the undersampling problem of the CSC through a locally optimized smoothing (see section 3a for details). The relative enhancement is found to be much stronger for the Gulf of Mexico than the Central MDR subregion, which can be explained by the more variable CSC. An optimal window length twice as large (48 vs. 24 days) is thus required for smoothing when averaged over the gridpoints within the subregion, leading to a more substantially modified CSCopt, and hence a potentially greater improvement in BSS. This explanation is also found for other subregions (not shown).

In terms of the NWP-based benchmark models, IDR-calibration helps increase S2STC BSSs by adding 3–6 % and 1–2 % for the Gulf of Mexico and the Central MDR, respectively, over the forecast weeks considered. For forecast week one, the S2STCcal model by far exceeds the CSCopt,

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0020.1.

but rapidly looses most of its skill over the first two forecast weeks, i.e. on the medium range, eventually leveling off thereafter on sub-seasonal timescales. While the CSCopt outperforms the S2STCcal from week three on in the Central MDR, the CSCopt takes the lead only beyond forecast week three in the Gulf of Mexico. Apart from these minor subregional differences, this considerable drop in model skill around week two to three is in accordance with previous findings for forecasts of basin-wide TC occurrence (Lee et al. 2018), highlighting the potential of climatological forecasts for sub-seasonal timescales.

Expanding the climatological model by including predictors generated from past data, the purely statistical approach improves the CSCopt skill for all five forecast weeks. While 3 % are added in the Gulf of Mexico at week one, improvements reduce to less than 0.7 % beyond medium range (Fig. 11a). In comparison, a maximum of 2 % is added to the CSCopt BSS in the Central MDR, but this level of improved skill drops to about 0.7 % only after week three (Fig. 11b). Considerable improvements can also be identified in subregions defined for the Caribbean Sea, and slightly north of the MDR (not shown), suggesting that sub-seasonal forecasts of weekly TC occurrence mainly for the MDR and adjacent subregions downstream can benefit from adding past data predictors.

Replacing the past data with the S2S ensemble mean and standard deviations for each predictor, the statistical–dynamical approach further raises the BSSs at all forecast weeks. The gain in skill is greatest for week one, and continuously decreases with longer lead times, except for minor sub-seasonal variations in the Central MDR. For the Gulf of Mexico, improvement in skill from the purely statistical to the statistical–dynamical approach is 4.5–6.5 (0.4–3.2) times greater on the medium (sub-seasonal) range than the improvement from the CSCopt to the purely statistical approach. In analogy, for the Central MDR, relative improvements appear to be 1.8–5.2 (0.2–3.8) times larger on the medium (sub-seasonal) range. Even though both logistic regression models are beaten by the S2STCcal model at week one, they outperform all benchmark models from week three (two) on in the Gulf of Mexico (Central MDR). Note that a simple approach to obtain at least equivalent skill for week one and two would be to include the S2STCcal forecasts as a predictor to the logistic regression models.

2) Comparison of different predictor sets

While a detailed analysis of predictor relevance is beyond the scope of this study, a simple approach elucidates the main sources for the predictive power of the statistical–dynamical model. Figure 12 provides insight into incremental improvements when successively including the predictor categories, summarized in Fig. 9, to the potential predictor set, from which the sequential predictor selection can choose the optimal subsets. Note that the inclusion of additional predictors may increase the degree of multicollinearity in the predictor set, which hence does not allow any conclusions to be drawn about potential deficiencies in predictive skill for the added category. In contrast, if an added predictor group improves skill, the improvement can clearly be attributed to the predictive skill inherent in the newly added group, regardless of whether multicollinearity is increased.

When first adding the GPI predictors to the CSCopt base predictor in the Gulf of Mexico, this already outperforms the purely statistical approach, which chooses from the full set of past data predictors, at all lead times (Fig. 12a). In the Central MDR, the oceanic predictors are included as the first group, which yields model skill that exceeds the purely statistical approach on the sub-seasonal timescale, and is almost comparable on the medium range (Fig. 12b). The majority of the sub-seasonal skill in the statistical-dynamical approach can be vastly attributed to the GPI (oceanic) predictor group for the Gulf of Mexico (Central MDR). The inclusion of the GPI predictors as the second group in the Central MDR leads to further improvements on the medium range, whereas skill increase by oceanic predictors added for the Gulf of Mexico is negligible. On the medium

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0020.1.

range, another substantial fraction of the skill results from adding information on tropical wave modes for both subregions.

7. Summary and conclusions

The main goal of the current study was to systematically validate and compare different types of models for probabilistic forecasts of weekly TC occurrence. By training a statistical model on predictors generated from NWP forecasts, a hybrid model was developed to leverage predictive skill, especially on sub-seasonal timescales. A variety of predictors was motivated and generated for various relevant predictor groups, covering climatological and oceanic information, environmental tropical influences, and their modulation by tropical wave modes, as well as extratropical influences. While in the statistical–dynamical approach, each predictor type was represented by the ensemble mean and standard deviation of the S2S ECMWF reforecasts, an analogous set of predictors was derived from ERA5 data to train a purely statistical approach. This approach, already applied in previous studies (Leroy and Wheeler 2008; Slade and Maloney 2013), served to address the question whether extracting predictors from NWP forecasts leads to better forecasting results. Optimal subsets of predictors have been determined for both approaches by running sequential predictor selections, before the statistical models were trained at every grid point separately. Climatological models (MSC, CSC, and CSCopt) and TCs tracked in the S2S model (S2STC, S2STCcal) served as benchmark. This set of model types has been validated for the Gulf of Mexico and Central MDR, with the following findings obtained:

- A simple CSC model already outperformed sub-seasonal predictions of a complex state-of-the-art NWP model, and therefore constituted a good base predictor for the development of statistical models. An optimized smoothing of the undersampled CSC led to considerable improvements, depending on the subregion.

- While the S2S ECMWF model predicted best at week one, it quickly dropped in skill thereafter due to the chaotic nature of the atmosphere blurring the valuable information contained in the initial conditions. This considerable sub-seasonal loss in skill confirms the findings of previous studies (e.g., Lee et al. 2018). The application of an IDR-calibration to the underforecasting S2STC model helped to raise skill at all lead times, but did not exceed the other approaches on sub-seasonal timescales.

- The purely statistical approach from (Leroy and Wheeler 2008; Slade and Maloney 2013), with logistic regression models trained on past data predictors, improved skill over the CSCopt model in both subregions out to week five. This corroborates the ability of statistical models to add to the sub-seasonal skill present in climatological models.

- With the statistical–dynamical approach, an even greater increase in model skill was found at all lead times considered, but especially on the medium range. Though this approach was still worse than the S2STCcal for week one, despite a significant increase in skill over the purely statistical approach, it outperformed all other models in the Gulf of Mexico (Central MDR) from week three (two) on. In the Gulf of Mexico, the sub-seasonal improvement from the purely statistical to the statistical–dynamical approach is 0.4–3.2 times larger as the one from the CSCopt to the purely statistical approach. The analogous for the Central MDR yields a factor for relative improvement of 0.2–3.8 . In view of the generally lower CSC skill in the Gulf of Mexico, such an improvement becomes even more remarkable, highlighting the value of this approach for subregions that are less subject to a seasonal cycle.

- When repeating the sequential predictor selection and training of the statistical–dynamical approach for CSCopt and GPI predictors only, the model was found to already perform better than the full predictor set provided to the purely statistical approach. A similar exceedance

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

in skill occurred in the Central MDR for the sub-seasonal lead times, when adding oceanic predictors at first. Furthermore, the majority of the additional sub-seasonal skill in the Gulf of Mexico (Central MDR) stemmed from the GPI (oceanic) predictors. Tropical wave modes were found to have their strongest skill contribution at medium range.

The systematic comparison of original and hybrid model types presented in this paper has demonstrated the great potential of statistical–dynamical modeling for a specific application of extreme events on the sub-seasonal forecast horizon. Exploiting S2S forecasts to develop such hybrid models proved to be the best strategy - at present - for probabilistic forecasting of subregional North Atlantic TC occurrence beyond week two, and might be a promising strategy for other (sub)basins and forecasting applications as well. Despite the identified improvements in forecast skill, predicting TC occurrence remains highly challenging, especially on sub-seasonal lead times. Given the hybrid model nature, there are two components, which - either alone or in combination - can be refined or even replaced with more sophisticated approaches to further leverage the predictive power. Regarding the underlying predictor set, more predictor types could be added, and/or their ensemble distribution could be represented in ways other than by the mean and standard deviation. Moreover, predictors could be constructed from other S2S models and/or with potentially higher spatio-temporal resolution. Alternatively, deep learning approaches could be used to automatize predictor extraction from full fields. As for the statistical component, the logistic regression model could be replaced by other parametric, but also nonparametric model approaches. Because TC occurrence is associated with several nonlinear processes (e.g., convection, tropical wave interaction, or extratropical Rossby-wave breaking), a follow-up study examines expected improvements due to deep learning architectures being able to model such nonlinearities.

Acknowledgments. The research leading to these results has been accomplished within projects C3 "Predictability of tropical and hybrid cyclones over the North Atlantic Ocean" and C5 "Dynamical feature-based ensemble postprocessing of wind gusts within European winter storms" of the Transregional Collaborative Research Center SFB/TRR 165 "Waves to Weather" funded by the German Science Foundation (DFG). This work is based on S2S data. S2S is a joint initiative of the World Weather Research Programme (WWRP) and the World Climate Research Programme (WCRP). The original S2S database is hosted at ECMWF as an extension of the TIGGE database. We thank Suzana J. Camargo, and two anonymous reviewers for their critical reviews and constructive suggestions. The authors further appreciate helpful discussions with Eva-Maria Walz and Tilmann Gneiting.

Data availability statement. The IBTrACS dataset used to calculate the target variable and climatological models is included in Knapp et al. (2010, 2018). Predictors for statistical models are generated from the S2S (Vitart et al. 2017) and ERA5 (Hersbach et al. 2020) datasets, respectively. The TCs tracked in S2S ECMWF reforecasts are openly available for download at `ftp://s2sidx:s2sidx@acquisition.ecmwf.int/TCYC`.

References

Akaike, H., 1974: A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **AC-19**, 716–723.

Belanger, J. I., J. A. Curry, and P. J. Webster, 2010: Predictability of North Atlantic tropical cyclone activity on intraseasonal time scales. *Mon. Wea. Rev.*, **138**, 4362–4374, doi:10.1175/2010MWR3460.1.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

Bluestein, H. B., 1993: *Synoptic-Dynamic Meteorology in Midlatitudes*, Vol. II. Oxford University Press, 608 pp.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Camargo, S. J., M. C. Wheeler, and A. H. Sobel, 2009: Diagnosis of the MJO modulation of tropical cyclogenesis using an empirical index. *J. Atmos. Sci.*, **66**, 3061–3074, doi:10.1175/2009JAS3101.1.

Camargo, S. J., and Coauthors, 2019: Tropical cyclone prediction on subseasonal time-scales. *Trop. Cyclone Res. Rev.*, **8**, 150–165, doi:10.1016/j.tcrr.2019.10.004.

Camp, J., and Coauthors, 2018: Skilful multiweek tropical cyclone prediction in ACCESS-S1 and the role of the MJO. *Quart. J. Roy. Meteor. Soc.*, **144**, 1337–1351, doi:10.1002/qj.3260.

Davis, C. A., and L. F. Bosart, 2003: Baroclinically induced tropical cyclogenesis. *Mon. Wea. Rev.*, **131**, 2730–2747, doi:10.1175/1520-0493(2003)131,2730:BITC.2.0.CO;2.

Davis, C. A., and L. F. Bosart, 2004: The TT problem: Forecasting the tropical transition of cyclones. *Bull. Amer. Meteor. Soc.*, **85**, 1657–1662, doi:10.1175/BAMS-85-11-1657.

Dimitriadis, T., T. Gneiting, and A. I. Jordan, 2021: Stable reliability diagrams for probabilistic classifiers. *Proc. Natl. Acad. Sci. USA*, **118**, e2016191 118, doi:10.1073/pnas.2016191118.

Dunion, J. P., and C. S. Velden, 2004: The impact of the Saharan air layer on Atlantic tropical cyclone activity. *Bull. Amer. Meteor. Soc.*, **85**, 353–366, doi:10.1175/BAMS-85-3-353.

Emanuel, K. A., 1986: An air-sea interaction theory for tropical cyclones. Part I: Steady-state maintenance. *J. Atmos. Sci.*, **43**, 585–605, doi:10.1175/1520-0469(1986)043<0585:AASITF>2.0.CO;2.

Emanuel, K. A., and D. S. Nolan, 2004: Tropical cyclone activity and the global climate system. *26th Conf. on Hurricanes and Tropical Meteorolgy*, Miami, FL, Amer. Meteor. Soc., 10A.2, [Available online at https://ams.confex.com/ams/26HURR/techprogram/paper_75463.htm.].

Fan, R.-E., K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, 2008: LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.*, **9**, 1871–1874, doi:10.1145/1390681.1442794.

Frank, W. M., and P. E. Roundy, 2006: The role of tropical waves in tropical cyclogenesis. *Mon. Wea. Rev.*, **134**, 2397–2417, doi:10.1175/MWR3204.1.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, doi:10.1198/016214506000001437.

Goldenberg, S. B., and L. J. Shapiro, 1996: Physical mechanisms for the association of El Niño and West African rainfall with Atlantic major hurricane activity. *J. Climate*, **9**, 1169–1187, doi:10.1175/1520-0442(1996)009<1169:PMFTAO>2.0.CO;2.

Gray, W. M., 1968: Global view of the origin of tropical disturbances and storms. *Mon. Wea. Rev.*, **96**, 669–700, doi:10.1175/1520-0493(1968)096<0669:GVOTOO>2.0.CO;2.

Gray, W. M., 1984: Atlantic seasonal hurricane frequency. Part I: El Niño and 30 mb quasi-biennial oscillation influences. *Mon. Wea. Rev.*, **112**, 1649–1668, doi:10.1175/1520-0493(1984)112<1649:ASHFPI>2.0.CO;2.

Gregory, P. A., J. Camp, K. Bigelow, and A. Brown, 2019: Sub-seasonal predictability of the 2017–2018 Southern Hemisphere tropical cyclone season. *Atmos. Sci. Lett.*, **20**, e886, doi:10.1002/asl.886.

Guo, C., G. Pleiss, Y. Sun, and K. Q. Weinberger, 2017: On calibration of modern neural networks. *International Conference on Machine Learning*, PMLR, 1321–1330.

Hastie, T., R. Tibshirani, and J. Friedman, 2009: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed., Springer Science & Business Media, 745 pp.

Henzi, A., J. F. Ziegel, and T. Gneiting, 2019: Isotonic distributional regression. arxiv.org, https://arxiv.org/abs/1909.03725.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, doi:10.1002/qj.3803.

Janiga, M. A., C. J. Schreck III, J. A. Ridout, M. Flatau, N. P. Barton, E. J. Metzger, and C. A. Reynolds, 2018: Subseasonal forecasts of convectively coupled equatorial waves and the MJO: Activity and predictive skill. *Mon. Wea. Rev.*, **146**, 2337–2360, doi:10.1175/MWR-D-17-0261.1.

Klotzbach, P., L.-P. Caron, and M. Bell, 2020: A statistical/dynamical model for North Atlantic seasonal hurricane prediction. *Geophys. Res. Lett.*, **47**, e2020GL089 357, doi:10.1029/2020GL089357.

Klotzbach, P., and Coauthors, 2019: Seasonal tropical cyclone forecasting. *Trop. Cyclone Res. Rev.*, **8**, 134–149, doi:10.1016/j.tcrr.2019.10.003.

Klotzbach, P. J., 2014: The Madden–Julian oscillation's impacts on worldwide tropical cyclone activity. *J. Climate*, **27**, 2317–2330, doi:10.1175/JCLI-D-13-00483.1.

Knapp, K. R., H. J. Diamond, J. P. Kossin, M. C. Kruk, and C. J. Schreck, 2018: International Best Track Archive for Climate Stewardship (IBTrACS) Project, Version 4, NA. NOAA National Centers for Environmental Information, accessed 14 April 2020, doi:10.25921/82ty-9e16.

Knapp, K. R., M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann, 2010: The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone best track data. *Bull. Amer. Meteor. Soc.*, **91**, 363–376, doi:10.1175/2009BAMS2755.1.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

Lee, C.-Y., S. J. Camargo, F. Vitart, A. H. Sobel, J. Camp, S. Wang, M. K. Tippett, and Q. Yang, 2020: Subseasonal predictions of tropical cyclone occurrence and ACE in the S2S dataset. *Wea. Forecasting*, **35**, 921–938, doi:10.1175/WAF-D-19-0217.1.

Lee, C. Y., S. J. Camargo, F. Vitart, A. H. Sobel, and M. K. Tippett, 2018: Subseasonal tropical cyclone genesis prediction and MJO in the S2S dataset. *Wea. Forecasting*, **33**, 967–988, doi: 10.1175/WAF-D-17-0165.1.

Leroy, A., and M. C. Wheeler, 2008: Statistical prediction of weekly tropical cyclone activity in the Southern Hemisphere. *Mon. Wea. Rev.*, **136**, 3637–3654, doi:10.1175/2008MWR2426.1.

Maier-Gerber, M., M. Riemer, A. H. Fink, P. Knippertz, E. Di Muzio, and R. McTaggart-Cowan, 2019: Tropical transition of Hurricane Chris (2012) over the North Atlantic ocean: A multiscale investigation of predictability. *Mon. Wea. Rev.*, **147**, 951–970, doi:10.1175/MWR-D-18-0188.1.

Maloney, E. D., and D. L. Hartmann, 2000: Modulation of hurricane activity in the Gulf of Mexico by the Madden-Julian oscillation. *Science*, **287**, 2002–2004, doi:10.1126/science.287.5460.2002.

Palmen, E., 1948: On the formation and structure of tropical hurricanes. *Geophysica*, **3**, 26–38.

Papin, P. P., L. F. Bosart, and R. D. Torn, 2020: A feature-based approach to classifying summertime potential vorticity streamers linked to Rossby wave breaking in the North Atlantic basin. *J. Climate*, **33**, 5953–5969, doi:10.1175/JCLI-D-19-0812.1.

Qian, Y., P.-C. Hsu, H. Murakami, B. Xiang, and L. You, 2020: A hybrid dynamical-statistical model for advancing subseasonal tropical cyclone prediction over the Western North Pacific. *Geophys. Res. Lett.*, **47**, e2020GL090 095, doi:10.1029/2020GL090095.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

Robertson, A. W., F. Vitart, and S. J. Camargo, 2020: Subseasonal to seasonal prediction of weather to climate with application to tropical cyclones. *J. Geophys. Res.: Atmos.*, **125**, e2018JD029 375, doi:10.1029/2018JD029375.

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2**, 191–201, doi: 10.1175/1520-0450(1963)002<0191:OSPF>2.0.CO;2.

Scheuerer, M., M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Wea. Rev.*, **148**, 3489–3506, doi:10.1175/MWR-D-20-0096.1.

Schreck III, C. J., J. Molinari, and A. Aiyyer, 2012: A global view of equatorial waves and tropical cyclogenesis. *Mon. Wea. Rev.*, **140**, 774–788, doi:10.1175/MWR-D-11-00110.1.

Schreck III, C. J., J. Molinari, and K. I. Mohr, 2011: Attributing tropical cyclogenesis to equatorial waves in the western North Pacific. *J. Atmos. Sci.*, **68**, 195–209, doi:10.1175/2010JAS3396.1.

Shapiro, L. J., 1982: Hurricane climatic fluctuations. Part II: Relation to large-scale circulation. *Mon. Wea. Rev.*, **110**, 1014–1023, doi:10.1175/1520-0493(1982)110<1014:HCFPIR>2.0.CO;2.

Slade, S. A., and E. D. Maloney, 2013: An intraseasonal prediction model of Atlantic and East Pacific tropical cyclone genesis. *Mon. Wea. Rev.*, **141**, 1925–1942, doi:10.1175/MWR-D-12-00268.1.

Thorncroft, C., and K. Hodges, 2001: African easterly wave variability and its relationship to Atlantic tropical cyclone activity. *J. Climate*, **14**, 1166–1179, doi:10.1175/1520-0442(2001)014<1166:AEWVAI>2.0.CO;2.

Van der Grijn, G., J. Paulsen, F. Lalaurette, and M. Leutbecher, 2005: Early medium-range forecasts of tropical cyclones. ECMWF Newsletter No. 102, ECMWF, Reading, United Kingdom, 7 pp.

Vitart, F., 2009: Impact of the Madden Julian Oscillation on tropical storms and risk of landfall in the ECMWF forecast system. *Geophys. Res. Lett.*, **36**, doi:10.1029/2009GL039089.

Vitart, F., A. Leroy, and M. C. Wheeler, 2010: A comparison of dynamical and statistical predictions of weekly tropical cyclone activity in the Southern Hemisphere. *Mon. Wea. Rev.*, **138**, 3671–3682, doi:10.1175/2010MWR3343.1.

Vitart, F., A. W. Robertson, and D. L. T. Anderson, 2012: Subseasonal to seasonal prediction project: Bridging the gap between weather and climate. *WMO Bull.*, **61**, 23–28.

Vitart, F., and T. N. Stockdale, 2001: Seasonal forecasting of tropical storms using coupled GCM integrations. *Mon. Wea. Rev.*, **129**, 2521–2537, doi:10.1175/1520-0493(2001)129<2521:SFOTSU>2.0.CO;2.

Vitart, F., and Coauthors, 2017: The subseasonal to seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, doi:10.1175/BAMS-D-16-0017.1.

Wang, Z., G. Zhang, T. J. Dunkerton, and F.-F. Jin, 2020: Summertime stationary waves integrate tropical and extratropical impacts on tropical cyclone activity. *Proc. Nat. Acad. Sci.*, **117**, 22 720–22 726, doi:10.1073/pnas.2010547117.

Wheeler, M., and G. N. Kiladis, 1999: Convectively coupled equatorial waves: Analysis of clouds and temperature in the wavenumber–frequency domain. *J. Atmos. Sci.*, **56**, 374–399, doi:10.1175/1520-0469(1999)056<0374:CCEWAO>2.0.CO;2.

Wheeler, M., and K. M. Weickmann, 2001: Real-time monitoring and prediction of modes of coherent synoptic to intraseasonal tropical variability. *Mon. Wea. Rev.*, **129**, 2677–2694, doi:10.1175/1520-0493(2001)129<2677:RTMAPO>2.0.CO;2.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932, doi:10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed., Elsevier, 676 pp.

Zhang, G., Z. Wang, T. J. Dunkerton, M. S. Peng, and G. Magnusdottir, 2016: Extratropical impacts on Atlantic tropical cyclone activity. *J. Atmos. Sci.*, **73**, 1401–1418, doi:10.1175/JAS-D-15-0154.1.

Zhang, G., Z. Wang, M. S. Peng, and G. Magnusdottir, 2017: Characteristics and impacts of extratropical Rossby wave breaking during the Atlantic hurricane season. *J. Climate*, **30**, 2363–2379, doi:10.1175/JCLI-D-16-0425.1.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

LIST OF TABLES

TABLE 1. Ranges of wave periods p and zonal wave numbers k used to filter for specific tropical wave modes. Definition is based on Janiga et al. (2018).

Wave type	Abbrev.	p [days]	k
Low-frequency	LF	> 100	$-10 : 10$
Madden-Julian Oscillation	MJO	$20 : 100$	$0 : 9$
Equatorial Rossby	ER	$10 : 100$	$-10 : -1$
Kelvin	Kelvin	$2.5 : 20$	$1 : 14$
Mixed Rossby-gravity/ Tropical depression	MRG/TD	$2.5 : 10$	$-20 : 0$

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

LIST OF FIGURES

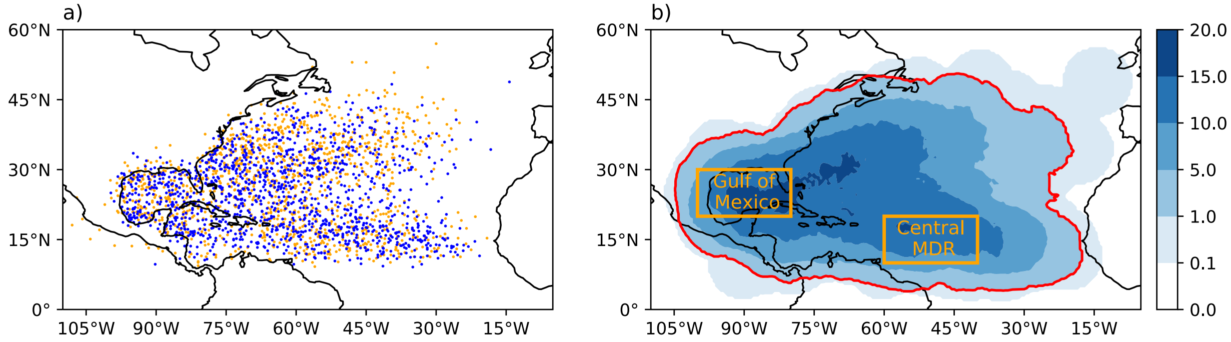


Fig. 1. (a) 1968-1997 (orange dots) and 1998-2017 (blue dots) IBTrACS tropical cyclone positions at 0000 UTC during the North Atlantic hurricane season (June-November) for intensities of at least tropical storm strength. (b) Relative frequency of TC occurrence (%) based on the definition in section 2a (see text for details). Note that interval boundaries are not equidistant. The red contour highlights the area where TCs occur at a rate of more than 1%. Orange boxes enclose the subregions used for model validation in section 5b.

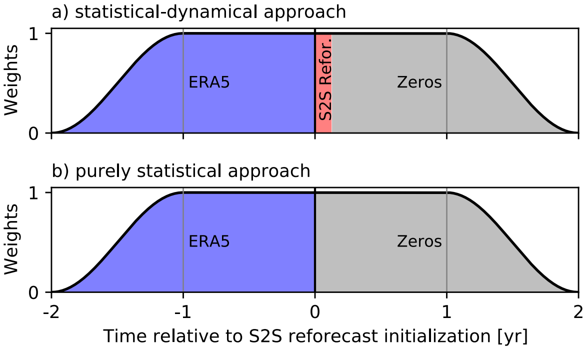


FIG. 2. Illustration of how time series used for real-time filtering are composed by different datasets, and how those are weighted over time in (a) the purely statistical, and (b) the statistical-dynamical approach, respectively.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

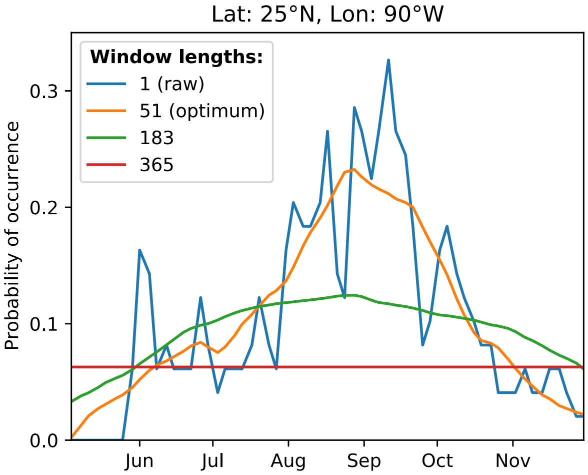


FIG. 3. Climatological seasonal cycle (CSC) at 25°N and 90°W exemplarily smoothed with uniform kernels of window length 1 (blue), 51 (orange), 183 (green), and 365 (red) days, respectively. At this grid point, the CSC smoothed with a window length of 51 days was identified to best correlate with the target variable.

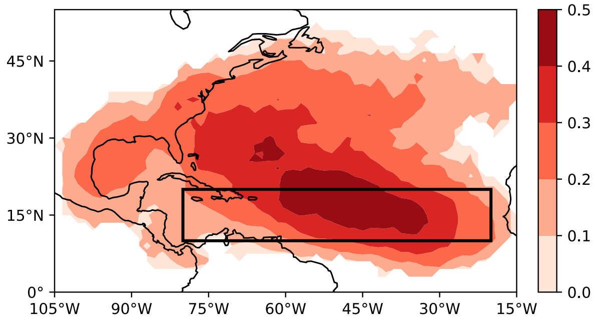


Fig. 4. Predictor–target Pearson correlation coefficient for the locally optimized climatology. Values are only displayed where correlations are statistically significant at a significance level of 5%. The thick black line highlights the MDR.

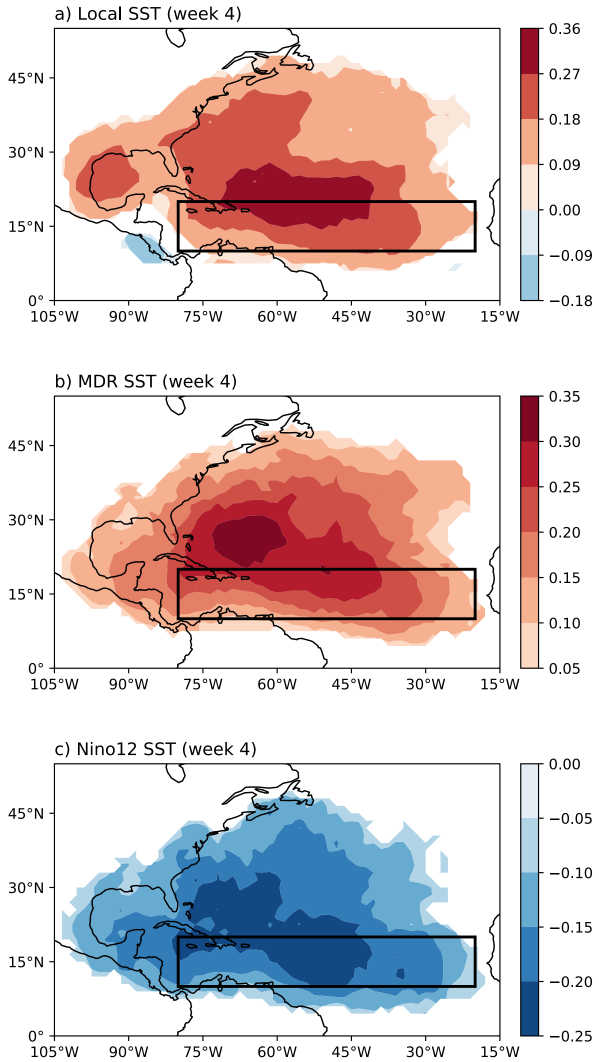
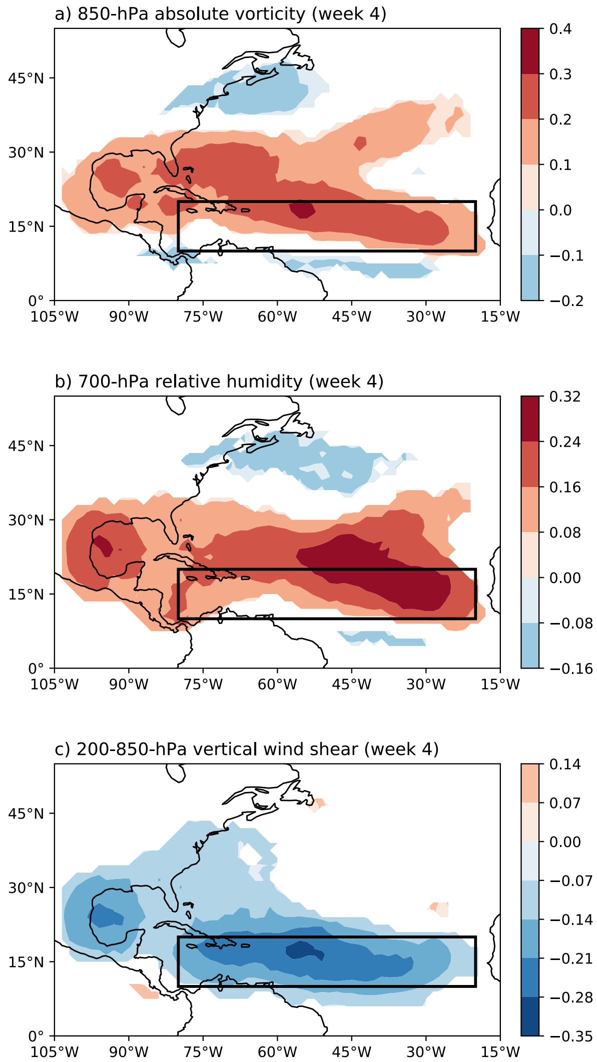


FIG. 5. As Fig. 4., but for ensemble mean (a) local SST, (b) MDR SST, and (c) Niño1+2 SST at week four. Note the varying ranges of different color bars.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.


a) 850-hPa absolute vorticity (week 4)

b) 700-hPa relative humidity (week 4)

c) 200-850-hPa vertical wind shear (week 4)

FIG. 6. As Fig. 4., but for ensemble mean (a) 850-hPa absolute vorticity, (b) 700-hPa relative humidity, and (c) 200–850-hPa wind shear at week four. Note the varying ranges of different color bars.

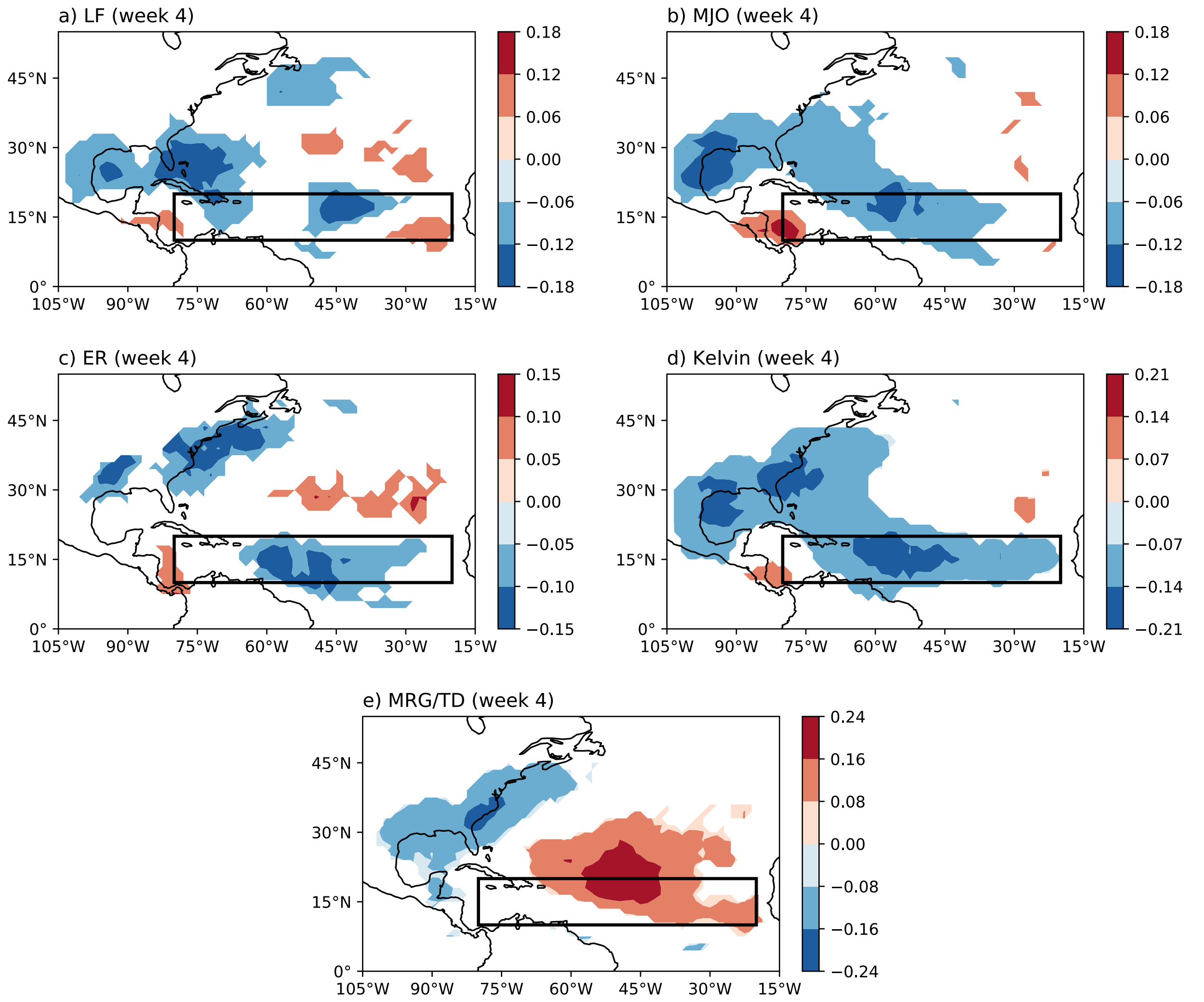


FIG. 7. As Fig. 4., but for ensemble mean 200-hPa divergence squared filtered for the wave types listed in Table 1 at week four. Note the varying ranges of different color bars.

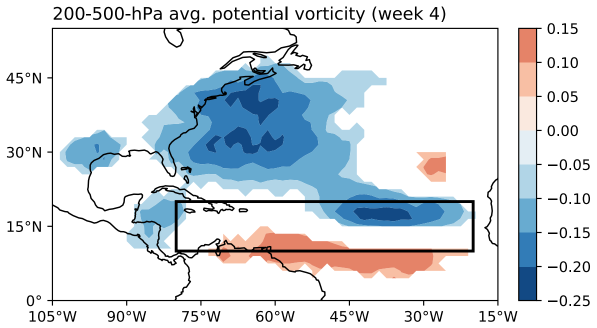
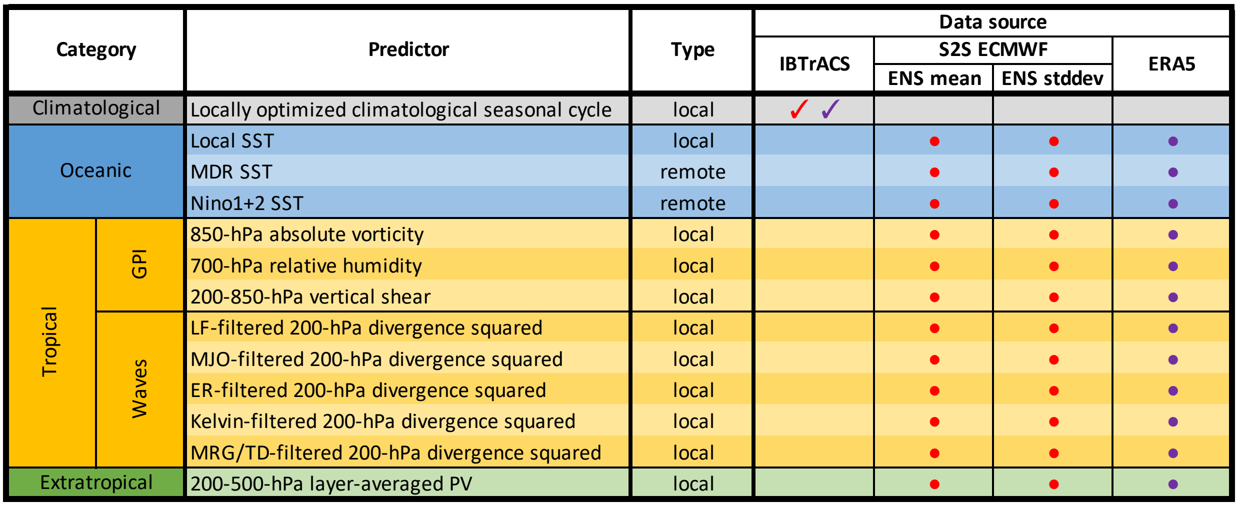


FIG. 8. As Fig. 4., but for ensemble mean 200–500-hPa layer-averaged PV at week four.



Category		Predictor	Type	Data source			
				IBTrACS	S2S ECMWF		ERA5
					ENS mean	ENS stddev	
Climatological		Locally optimized climatological seasonal cycle	local	✓ ✓			
Oceanic		Local SST	local		●	●	●
		MDR SST	remote		●	●	●
		Nino1+2 SST	remote		●	●	●
Tropical	GPI	850-hPa absolute vorticity	local		●	●	●
		700-hPa relative humidity	local		●	●	●
		200-850-hPa vertical shear	local		●	●	●
	Waves	LF-filtered 200-hPa divergence squared	local		●	●	●
		MJO-filtered 200-hPa divergence squared	local		●	●	●
		ER-filtered 200-hPa divergence squared	local		●	●	●
		Kelvin-filtered 200-hPa divergence squared	local		●	●	●
		MRG/TD-filtered 200-hPa divergence squared	local		●	●	●
Extratropical		200-500-hPa layer-averaged PV	local		●	●	●

FIG. 9. Overview schematic showing the set of predictors, from which the stepwise predictor selection chooses an optimal predictor set for every gridpoint and forecast week. Red and purple symbols indicate predictors of the statistical-dynamical and the purely statistical models, respectively. While dots denote predictors that can be chosen by the sequential predictor selection, ticks signify fixed predictors.

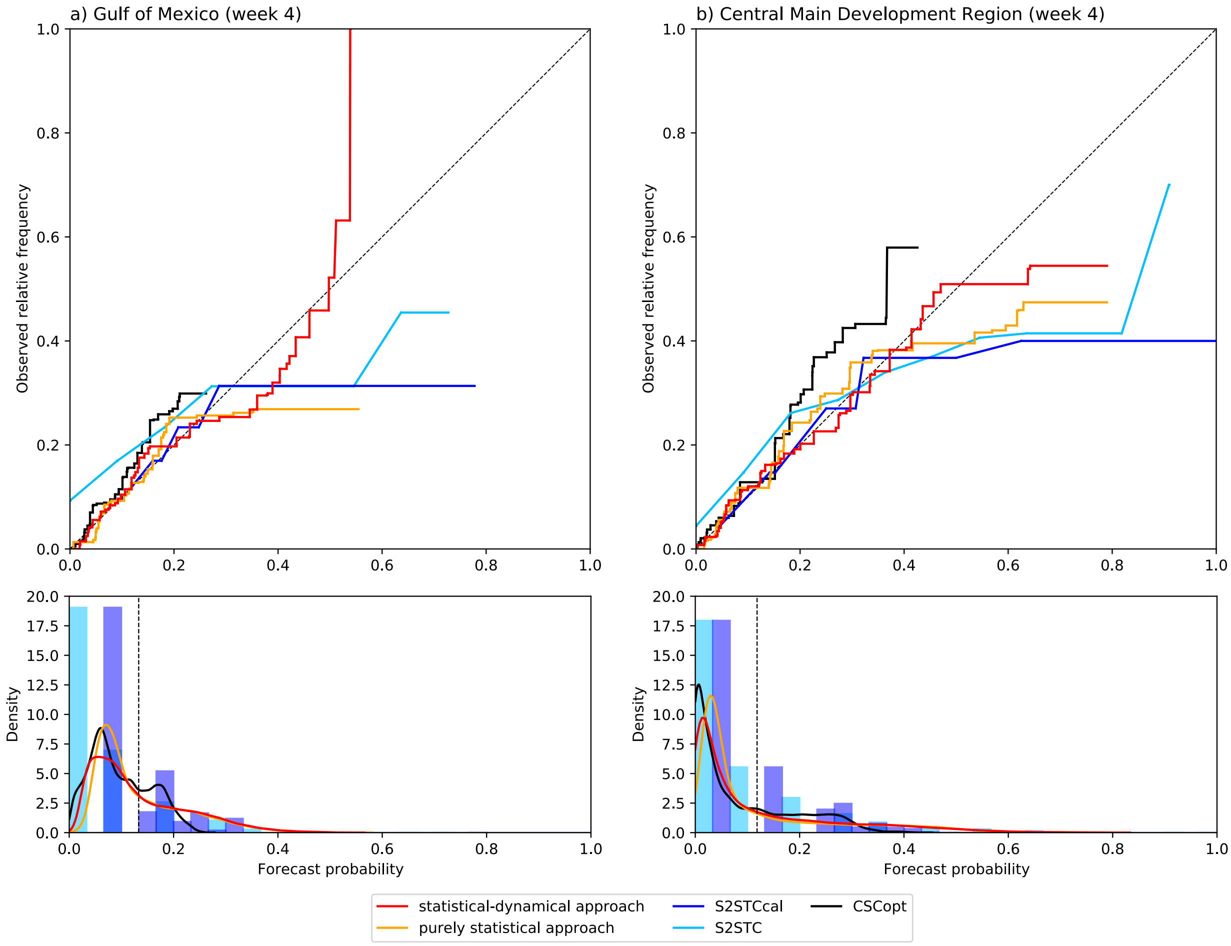


FIG. 10. CORP reliability diagram for (a) Gulf of Mexico and (b) Central MDR week-four forecasts, respectively. While forecast probabilities distributions are visualized by means of histograms for the S2STC and S2STCcal models, a kernel density estimation is applied to generated continuous curves for the other models. The dashed vertical line indicates the mean relative frequency of the target variable.

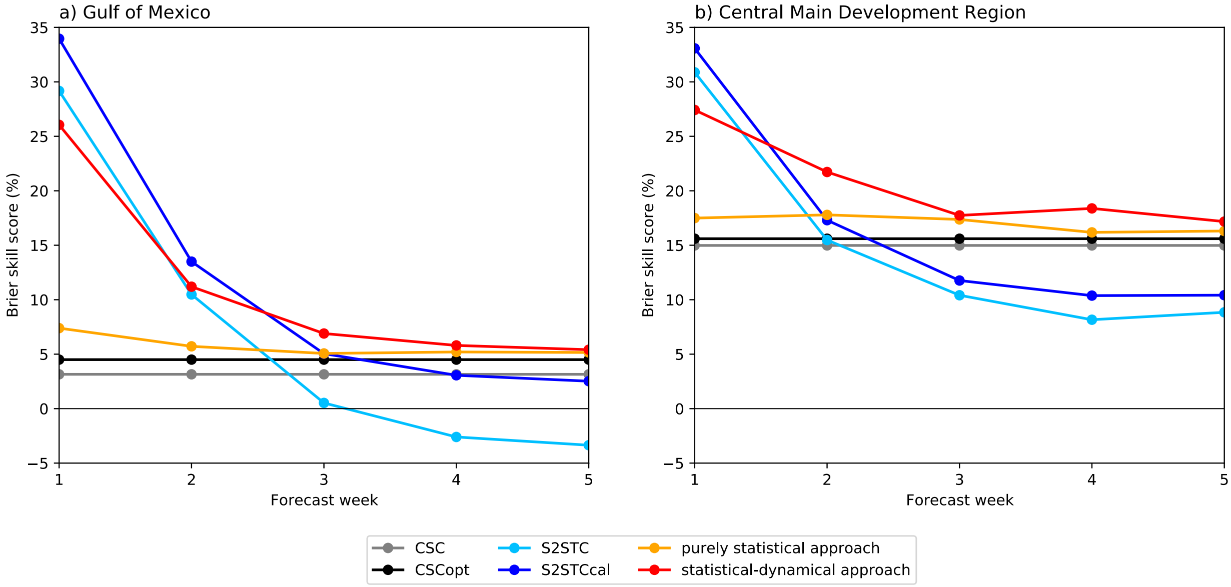


FIG. 11. Brier skill score (BSS; %) as a function of forecast week for the CSC (gray), CSCopt (black), S2STC (lightblue), S2STCcal (darkblue), purely statistical (orange), and statistical–dynamical (red) models, respectively, relative to the MSC and validated in the Gulf of Mexico (left) and Central MDR (right) subregions.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0020.1.

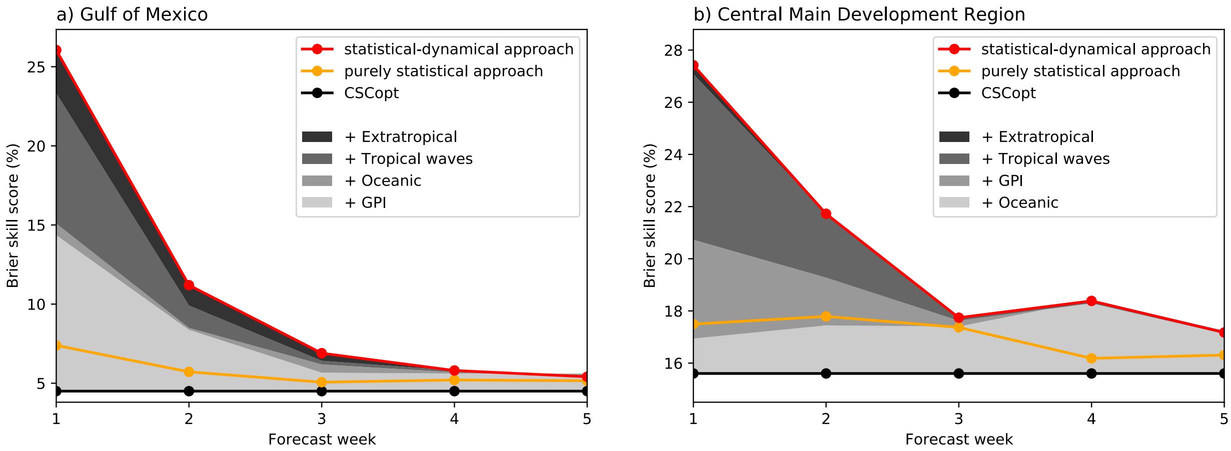


FIG. 12. Similar to Fig. 11, but only the CSCopt, purely statistical, and statistical–dynamical models are shown in black, orange, and red lines, respectively, relative to the MSC. For the latter, improvements in BSS, when successively including the predictor categories (see Fig. 9) to the sequential predictor selection, are illustrated by grayish shadings. Note that (a) and (b) differ in that the order of inclusion is reversed for the first two predictor groups. For better visualization, different y-axis ranges were used.