Subseasonal tropical cyclone activity over the North Atlantic:

A systematic comparison of different forecast approaches

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

(Dr. rer. nat.)

von der KIT-Fakultät für Physik des

Karlsruher Instituts für Technologie (KIT)

angenommene

DISSERTATION

von

M. Sc. Michael Maier-Gerber

Tag der mündlichen Prüfung:	27.01.2023
Referent:	Prof. Dr. Andreas H. Fink
Korreferenten:	Prof. Dr. Peter Knippertz
	PD Dr. habil. Michael Riemer (JGU Mainz)

Abstract

The devastating winds, torrential rainfalls, and storm surges associated with tropical cyclones (TCs) frequently claim numerous fatalities and inflict extensive and costly damages. Planning evacuations and taking precautionary measures requires accurate forecasts of TC activity with as much lead time as possible. Past efforts have mostly focused on predicting either individual TCs several days ahead or the activity of an entire season. This separation is primarily due to the subseasonal predictability gap (beyond 2 weeks but less than 3 months), and is also reflected in the fact that different modeling approaches are predominantly used for each of the two forecast ranges. Medium-range predictions (up to 2 weeks) are heavily based on numerical weather prediction (NWP; also referred to as 'dynamical') models, whereas statistical models are usually trained to issue seasonal (3-6 months) outlooks. While previous research on subseasonal TC activity has mostly focused on either the validation of NWP models, or the development of statistical models trained on past data, the present thesis combines both approaches to a statistical–dynamical (or 'hybrid') model for probabilistic forecasts in the North Atlantic basin. This dissertation aims to identify and examine NWP-based predictors relevant for subseasonal TC activity forecasting, to develop and validate a statistical-dynamical forecasting model, and to systematically compare its predictive performance with a variety of distinct forecasting approaches.

Although state-of-the-art NWP models were shown to lack predictive skill with regard to subseasonal weekly TC activity, they may predict the environmental conditions sufficiently well to generate predictors for a statistical model. Therefore, based on a literature review and physical considerations, an extensive predictor set is generated including predictor groups representing climatological and seasonal information, oceanic, and tropical conditions, tropical wave modes, as well as extratropical influences. The assumption that these predictors provide exploitable predictive skill at subseasonal lead times is considered to be valid, as for each week-four predictor significant correlations with TC occurrence are found between 0.1 and 0.5.

A statistical-dynamical (or hybrid) model is developed for predicting TC occurrence probability and the predictive distribution of accumulated cyclone energy (ACE) for lead

times up to week five, using logistic regression and a two-part model, respectively. To contrast differences between subregions, separate models are trained and validated for the Gulf of Mexico and the central Main Development Region (MDR), respectively. For each forecast week and subregion, an automated procedure selects only relevant predictors from the predictor pool, before the statistical model component is applied in forecast mode at every grid point separately. Even though regularisation is imposed to prevent the model from over- and underfitting, most predictors are still selected during this procedure, attesting their utility for the hybrid approach. A variety of original and optimized models, including climatological models, purely dynamical models, and purely statistical models provide a comprehensive set of benchmarks. Beyond the purpose of facilitating a strong and thus honest competition for model validation, it also allows to assess predictive performance for a hierarchy of modelling approaches.

This variety of models is systematically cross-validated on the 1979–2018 period for predictions in the Gulf of Mexico and central MDR subregions, respectively. The verification of probabilistic forecasts combines established tools with newly developed techniques to assess the calibration of models, their potential and actual predictive skills, and the expected long-term costs for a user when taking action based on each model. The climatological and NWP-based models are found to systematically underforecast both target variables, which can be corrected in the latter case of model type by statistical post-processing. In contrast, the purely statistical and the statistical-dynamical models produce overall well calibrated forecasts. The NWP models perform best at week one but quickly lose skill within the first two weeks due to the chaotic nature of the atmosphere blurring the valuable information contained in the initial conditions. Even in case of recalibration, the NWP models are outperformed by the climatological models on subseasonal lead times. Seasonal variations reflected in a climatological model are particularly useful in the central MDR, which is subject to a more pronounced seasonal cycle. In contrast, an optimization of how much information from adjacent days in the year is used to compile a probabilistic climatological distribution seems to pay off most in the Gulf of Mexico, because more instances are required to obtain a robust distribution. The purely statistical models increase skill over the climatological models only slightly, suggesting that past information does not contain much of exploitable predictive skill. The statistical-dynamical approach achieves considerable skill improvements in predicting TC occurrence up to week five for both subregions. The vast majority of the additional subseasonal skill in the hybrid model, relative to the climatological model, can be attributed to the tropical conditions in the Gulf of Mexico, and to the oceanic conditions in the central MDR. For the predictive distribution of ACE, skill improvements are qualitatively similar but disappear beyond week three in the Gulf of Mexico. Training the ACE model while withholding the climatological base predictor demonstrates the

occurrence factor to be of much greater importance than the intensity factor. Applying a cost-loss decision model to the TC occurrence predictions broadly suggests that, to reduce overall economic costs, the most valuable information is provided by the NWP model on the medium range, and by the statistical-dynamical model on the subseasonal range.

Kurzfassung

Die zerstörerischen Winde, sintflutartigen Regenfälle und Sturmfluten, die mit tropischen Wirbelstürmen (TCs) einhergehen, fordern häufig zahlreiche Todesopfer und verursachen großflächige und kostenintensive Schäden. Die Planung von Evakuierungen und Vorsichtsmaßnahmen erfordert genaue Vorhersagen der TC-Aktivität mit einer möglichst großen Vorlaufzeit. Vergangene Bemühungen haben zumeist darauf abgezielt entweder einzelne TCs mit einer Vorlaufzeit von mehreren Tagen oder die Aktivität einer ganzen Saison vorherzusagen. Diese Trennung ist in erster Linie auf die subsaisonale Vorhersagbarkeitslücke (mehr als 2 Wochen aber weniger als 3 Monate) zurückzuführen, und spiegelt sich auch in der Tatsache wider, dass für jeden der beiden Vorhersagebereiche vorwiegend unterschiedliche Modellierungs-ansätze zur Anwendung kommen. Mittelfristige Vorhersagen (bis zu 2 Wochen) basieren in hohem Maße auf numerischen Wettervorhersagemodellen (NWP-Modell; auch als 'dynamisches Modell' bezeichnet), während statistische Modelle in der Regel für saisonale Vorhersagen (3-6 Monate) trainiert werden. Während frühere Forschung zur sub-saisonalen TC-Aktivität sich überweigend entweder auf die Validierung von NWP-Modellen oder auf die Entwicklung statistischer Modelle, trainiert auf vergangenen Daten, konzentriert hat, kombiniert die vorliegende Arbeit beide Ansätze zu einem statistisch-dynamischen (oder hybriden) Modell für probabilistische Vorhersagen im nordatlantischen Ozean. Diese Dissertation zielt darauf ab, NWP-basierte Prädiktoren, die für die Vorhersage subsaisonaler TC-Aktivität relevant sind, zu identifizieren und zu bewerten, ein statistisch-dynamisches Vorhersagemodell zu entwickeln und zu validieren und dessen Vorhersageleistung systematisch mit einer Vielzahl von unterschiedlichen Vorhersageansätzen zu vergleichen.

Obwohl es modernen NWP-Modellen nachweislich an der Fähigkeit zur Vorhersage subsaisonaler, wöchentlicher TC-Aktivität mangelt, könnten sie in der Lage sein die Umgebungsbedingungen ausreichend gut vorherzusagen, um daraus Prädiktoren für ein statistisches Modell zu erzeugen. Daher wird auf Grundlage einer Literaturrecherche und physikalischen Überlegungen ein umfangreicher Satz and Prädiktoren generiert, welcher Prädiktorgruppen umfasst, die klimatologische und saisonale Informationen, ozeanische und tropische Bedingungen, tropische Wellenmoden sowie außertropische Einflüsse

repräsentieren. Die Annahme, dass diese Prädiktoren für subsaisonale Vorlaufzeiten nutzbare Vorhersagefähigkeiten aufweisen, wird als zutreffend erachtet, da für jeden Prädiktor der Vorhersagewoche vier signifikante Korrelationen mit dem Auftreten von TCs zwischen 0.1 und 0.5 zu finden sind.

Ein statistisch-dynamisches (oder Hybrid-) Modell wird für die Vorhersage der Auftretenswahrscheinlichkeit von TCs sowie für die Accumulated Cyclone Energy (ACE) mit Vorlaufzeiten bis zur Woche fünf entwickelt, wobei eine logistische Regression bzw. ein zweiteiliges Modell verwendet wird. Um Unterschiede zwischen den Teilregionen vergleichen zu können, werden separate Modelle für den Golf von Mexiko bzw. die zentrale Main Development Region (MDR) trainiert und validiert. Für jede Vorhersagewoche und Teilregion wählt ein automatisiertes Verfahren lediglich die relevanten Prädiktoren aus dem Prädiktorenpool aus, bevor die statistische Modellkomponente im Vorhersagemodus auf jeden einzelnen Gitterpunkt separat angewendet wird. Obwohl eine Regularisierung vorgenommen wird, um eine Über- oder Unteranpassung des Modells zu verhindern, werden dennoch die meisten Prädiktoren während dieses Verfahrens ausgewählt, was deren Nützlichkeit für den hybriden Ansatz belegt. Eine Vielfalt an originären und weiterverbesserten Modellen, darunter klimatologische Modelle, rein dynamische Modelle und rein statistische Modelle, bieten eine umfassende Auswahl an Benchmarks. Neben dem Zweck, einen starken und somit fairen Wettbewerb während der Modellvalidierung zu ermöglichen, erlaubt dies auch die Bewertung der Vorhersageleistung einer Hierarchie an Modellierungsansätzen.

Diese Vielfalt an Modellen wird systematisch für Vorhersagen in den Teilregionen Golf von Mexiko bzw. zentrale MDR über den Zeitraum 1979–2018 kreuzvalidiert. Die Verifikation der probabilistischen Vorhersagen kombiniert etablierte Methoden mit neu entwickelten Techniken, um die Kalibrierung der Modelle und ihre potenzielle und tatsächliche Vorhersagefähigkeit auszuwerten, sowie die langfristig zu erwartenden Kosten für einen Nutzer abzuschätzen, der auf Grundlage des jeweiligen Modelles Maßnahmen ergreifen würde. Es zeigt sich, dass die klimatologischen und NWP-basierten Modelle beide Zielvariablen systematisch unterschätzen, was im letzteren Fall an Modelltyp durch statistisches Postprocessing korrigiert werden kann. Im Gegensatz dazu liefern die rein statistischen und die statistisch-dynamischen Modelle insgesamt gut kalibrierte Vorhersagen. Die NWP-Modelle schneiden in der ersten Woche am besten ab, verlieren aber innerhalb der ersten zwei Wochen schnell ihre Vorhersagefähigkeit, da die chaotische Natur der Atmosphäre die in den Anfangsbedingungen enthaltenen wertvollen Informationen unscharf werden lässt. Selbst im Falle einer Rekalibrierung werden die NWP-Modelle von den klimatologischen Modellen für subsaisonale Vorlaufzeiten übertroffen. Saisonale Schwankungen, die sich in einem klimatologischen Modell widerspiegeln, sind besonders in der zentralen MDR nützlich, die einem ausgeprägteren saisonalen Zyk-

lus unterliegt. Im Gegensatz dazu scheint sich eine Optimierung, wie viele Informationen von benachbarten Tagen im Jahr verwendet werden, um eine probabilistische klimatologische Verteilung zu erstellen, im Golf von Mexiko am meisten auszuzahlen, da mehr Instanzen erforderlich sind, um eine robuste Verteilung zu erhalten. Die rein statistischen Modelle erhöhen die Vorhersagefähigkeit gegenüber den klimatologischen Modellen nur geringfügig, was darauf hindeutet, dass Informationen aus der Vergangenheit nicht viel an nutzbarer Vorhersagefähigkeit enthalten. Der statistisch-dynamische Ansatz erzielt erhebliche Verbesserungen bei der Vorhersage des Auftretens von TCs bis zur Woche fünf für beide Teilregionen. Der überwiegende Teil der zusätzlichen subsaisonalen Fähigkeit des Hybridmodells im Vergleich zum klimatologischen Modell kann den tropischen Bedingungen im Golf von Mexiko bzw. den ozeanischen Bedinungen in der zentralen MDR zugeschrieben werden. Die Vorhersageverbesserungen für die Verteilung von ACE sind qualitativ ähnlich, verschwinden aber nach der dritten Woche im Golf von Mexiko. Das Training des ACE-Modells unter Zurückhaltung des klimatologischen Basisprädiktors zeigt, dass der Auftretensfaktor von wesentlich größerer Bedeutung ist als der Intensitätsfaktor. Die Anwendung eines Kosten-Verlust-Entscheidungsmodells auf die Vorhersagen des Auftretens von TCs deutet weitgehend darauf hin, dass, zur Senkung der wirtschaftlichen Gesamtkosten, die wertvollsten Informationen das NWP-Modell für den mittelfristigen und das statistisch-dynamische Modell für den subsaisonalen Bereich liefert.

Preface

The PhD candidate confirms that the research presented in this thesis contains significant scientific contributions by himself. This thesis reuses material from the following publication:

> Maier-Gerber, M., A. H. Fink, M. Riemer, E. Schoemer, C. Fischer, and B. Schulz, 2021: Statistical-Dynamical Forecasting of Subseasonal North Atlantic Tropical Cyclone Occurrence. *Weather and Forecasting*, **36 (6)**, 2127-2142, doi:10.1175/WAF-D-21-0020.1.

The abstract and Chapters 1, 2, 3, 4, 5, 6, 8, and 9 reuse material from Maier-Gerber et al. (2021). © 2021, American Meteorological Society. Used with permission.

The research leading to these results has been conducted within project C3 ”Predictability of tropical and hybrid cyclones over the North Atlantic Ocean” of the Transregional Collaborative Research Center SFB/TRR 165 ”Waves to Weather”, funded by the German Research Foundation (DFG). The research proposal of this project was written by Andreas H. Fink, Michael Riemer, and Elmar Schoemer, with significant contributions from the candidate. Analyses in Maier-Gerber et al. (2021) were solely performed by the candidate, who also wrote the text with advice from Andreas H. Fink and Michael Riemer and comments from all co-authors during the manuscript preparation. Chapter 7 was solely performed by the candidate, who also wrote the text with comments from Andreas H. Fink and Benedikt Schulz.

Contents

1. Introduction

Tropical cyclones (TCs) are among the most impressive weather phenomena on our planet, as they are characterized by extreme deviations from the mean atmospheric state. The characteristic deviations of this phenomenon manifest in various meteorological and oceanic variables (e.g., in atmospheric pressure, wind speed, precipitation, or ocean wave height), occasionally leading to new record values. In part, it is certainly due to this fascinating extreme nature that TCs generally receive such great attention in the public and the media. However, interest in TCs is many times greater in countries whose populations, economies and infrastructures are regularly affected by the associated hazards and damage. In the US, for instance, TCs were accountable for most fatalities and economic losses associated with billion-dollar catastrophic events between 1980 and 2022 among all types of natural disasters (NOAA NCEI, 2022). The $6,864$ deaths and total costs of $\$1,194.4$ billion (consumer price index adjusted) attributed to the impact of TCs during this 43-years period give every reason to further invest in the development of early warning systems, to expand information campaigns for populations along affected coastlines, and to continuously evaluate and improve evacuation concepts of decision-makers.

The main contribution of the meteorological discipline to mitigate TC-related impacts concerns the early warning systems, with the aim of providing and updating forecasts, including their uncertainty, as early as possible. All efforts undertaken can be broadly divided into two fields of activity. On the one hand, research on TCs and the associated limits of predictability is essential to gain a better understanding of this phenomenon as well as to provide a sound knowledge base for the development and implementation of forecast models. On the other hand, weather services operationally run, evaluate and continuously improve their models, to eventually disseminate various forecast products for a wide range of users in graphical and textual form. Forecasts for individual TCs are typically issued for the next few days (e.g., up to five days ahead by the National Hurricane Center) and heavily based on the output from numerical weather prediction (NWP; alternatively referred to as *dynamical*) models. In contrast, seasonally integrated TC activity is predicted by different institutional groups and mostly builds on statistical models (Camargo et al., 2007). This coexistence of distinct lead times and model approaches is

due to the subseasonal predictability gap (Vitart et al., 2012; Robertson et al., 2020), a period of lead times – beyond 2 weeks but less than 3 months – at the lower end of the transition between weather and climate time scales for which predictability is generally reduced.

Because weather phenomena such as TCs stand out for their extreme deviations, they are often associated with low predictability, so that their prediction on subseasonal timescales poses an even greater challenge. Several studies have evaluated NWP models in terms of their subseasonal predictive skill for various TC activity measures (e.g., Lee et al., 2018, 2020; Gregory et al., 2019). Examining predictions for TC occurrence from different operational forecast centers, Lee et al. (2018) conclude that dynamical models generally have little to zero skill from week two on in all basins relative to climatological forecasts. For the North Atlantic, they even state that actual and potential model skill is very close, suggesting that hardly any improvement can be achieved with current NWP models. However, the growing understanding of various modes of subseasonal and interannual variability, such as the Madden-Julian Oscillation (MJO; Madden and Julian, 1971, 1972) and the El Niño Southern Oscillation (ENSO; Wang et al., 2017), and their potential role as sources for subseasonal TC predictability has led to an increased research focus (Camargo et al., 2019). With dynamical models nowadays often integrated to subseasonal or seasonal forecast horizons, these atmospheric modes of variability have been shown to influence subseasonal forecasts for TC activity in many oceans (e.g., Vitart, 2009; Belanger et al., 2010; Camp et al., 2018).

Even though partly living on shorter time scales, tropical wave modes and extratropical Rossby wave breaking carry longer-lived predictive signals that can be exploited as additional sources of subseasonal predictability (Janiga et al., 2018; Frank and Roundy, 2006; Papin et al., 2020). All this suggests that NWP models, although lacking skill when directly predicting subseasonal TC activity, may be able to predict the environmental conditions favourable for TC genesis with sufficient skill, so that predictors can be generated and fed into statistical models. The present study intends to assess the expected predictive value of such a combined statistical-dynamical (or hybrid) forecasting approach for subregional TC activity in the North Atlantic. For this purpose, a variety of climatological, oceanic, tropical, and extratropical predictors is generated, which are known to precondition and modulate environments to become favourable for TCs. Models are developed for two target variables – one for TC occurrence and the second for an intensity-related measure of TC activity. A well-founded assessment of the hybrid model performance requires a validation in comparison to a number of distinct benchmark models, including climatological, purely dynamical, and purely statistical approaches. Accordingly, this dissertation focuses on the following three overarching research aims:

- to identify and examine NWP-based predictors relevant for subseasonal TC activity forecasting,
- to develop and validate a statistical-dynamical forecasting model,
- to systematically compare its predictive performance with a variety of distinct forecasting approaches.

In the beginning of this thesis, Chapter 2 provides the theoretical background on TCs, on the subseasonal predictability gap, as well as on the basic concepts of dynamical, statistical, and statistical-dynamical modelling, along with current applications in TC forecasting. Based on the overarching research aims stated above, Chapter 3 then motivates and formulates a set of concrete research questions that shall be answered. Following this, Chapter 4 describes the data and methods used to develop and validate models, before a pool of relevant predictors for the development of the statistical models is elaborated on the basis of existing literature and physical considerations in Chapter 5. Chapter 6 provides a detailed description of generated benchmark models, and statistical models developed for TC occurrence prediction, followed by a systematic validation. Chapter 7 follows the same structure for the model development and validation for forecasting the intensity-related measure of TC activity. At the end of this thesis, the lessons learned and potential for future work is discussed in the conclusions in Chapter 8 and in the outlook in Chapter 9, respectively.

2. Theoretical background

2.1 Tropical cyclones

2.1.1 Characteristics, impact, and variability

In the North Atlantic and East Pacific, TCs are also referred to as 'hurricanes', while other regional names are used in other oceans. TCs are axisymmetric low-pressure systems with exceptionally high radial gradients, reaching minimum central pressure values up to less than 900 hPa. To a first approximation, their horizontal (or primary) circulation results from gradient wind balance between pressure gradient force, horizontal centrifugal force, and Coriolis force. Friction in the boundary layer reduces the latter two forces and creates an inflow layer, in which air parcels spin up along cyclonic trajectories due to conservation of absolute angular momentum. The resulting convergence leads to updrafts in the inner-core region, followed by anticyclonic outflow in the upper troposphere, and weak compensating subsidence, completing the vertical (or secondary) circulation. Because TCs occur over relatively warm oceans, the inflowing air parcels acquire high entropy from surface heat and moisture fluxes before rising in deep cumulus convection organized in an annular eyewall typically few decakilometers in diameter. While adiabatically expanding in the updrafts, the air parcels release their latent heat, fostering the development and intensification of an upper-level warm-core. This thermal characteristic in combination with the non-frontal structure set TCs apart from their extratropical and subtropical counterparts, which feature an upper-level cold-core and/or exhibit some asymmetric frontal structure.

The heating of the core reduces the central pressure, which in turn increases the pressure gradients, further driving the circulation. Therefore, the minimum central pressure at sea-level usually serves as a good measure for TC intensity. Another commonly used indicator is maximum sustained winds at 10-m height, which typically reaches peak values at the top of the atmospheric boundary layer within the eyewall. Unlike wind gust variables, this measure does not express instantaneous values but is averaged over a certain time period. While the World Meteorological Organization (WMO) recommends to use a 10-minute averaging period (Harper et al., 2010), it is common practice at the US

National Weather Service to use a 1-minute period for the North Atlantic and Eastern Pacific. Based on this deviating convention, a TC is called a 'tropical depression' if its intensity does not exceed 33 kt, whereas it is rated as a 'tropical storm' for maximum sustained winds between 34 kt and 63 kt. For higher intensities, the Saffir–Simpson hurricane wind scale (SSHWS; Simpson, 1974) distinguishes 5 categories, beginning with Category 1 at an exceedance of 64 kt, followed by Category 2, 3, 4, and 5 when 83, 96, 113, 137 kt are reached, respectively. A TC classified as Category 3 or higher is also referred to as a 'major hurricane'.

When considering the overall potential impact of a TC, however, it is not only maximum wind speed that matters, but also the spatial extent of the wind field. Using an empirical regression model for economic loss and keeping maximum winds constant, Zhai and Jiang (2014) found that costs associated with the landfall of Hurricane Sandy in 2012 would have been lower by a factor of 20 if the TC had been three times smaller in radius, and thus of average extent. Moreover, a very intense TC with highly localized extreme winds may cause less overall damage than a weaker TC with a broader wind field. The radius of a TC can vary largely from case to case, as shown by the record values for 34-kt winds, which range from about 20 km for Tropical Storm Marco in 2008 (Knapp et al., 2018) to over 1000 km for Typhoon Tip in 1979 (Dunnavan and Diercks, 1980). Therefore, various quantities, such as the radius of maximum winds, or quadrant-specific values for the radial extent of certain wind speeds (e.g., 34, 50, or 64 kt) were specified to address this problem. Beyond all considerations of the spatial extent of a TC, a frequently raised criticism of the SSHWS definition is that it merely considers intensity and completely disregards other relevant aspects of impact, such as heavy rainfall, storm surge, and cyclone movement. For example, the landfall of Hurricane Harvey in 2017 led to enormous inland flooding in Baytown, Texas, as the stationary TC brought total rainfall amounts of more than 1000 mm accumulated over 3 days (Van Oldenborgh et al., 2017).

Even though all considerations about the intensity of TCs, their spatial extent, and the relevance of different impact-related variables have their justification, they presuppose the existence of a TC. On average, around 80 TCs occur per year in all global basins combined. Over the past decades, numerous field campaigns and idealized simulations have led to different theoretical concepts and a gernally growing understanding of TC genesis (Tang et al., 2020), but a holistic theory that could explain this global frequency of TC occurrence does not exist to date (Sobel et al., 2021). Because of the complexity of their structure and circulation, and the interplay of the associated dynamic and thermodynamic processes described above, TCs are not a mere result of environmental conditions. In fact, there is a wide range of potential interactions with different types of atmospheric and oceanic features during the various stages of a TC's life cycle. If not isolated from the

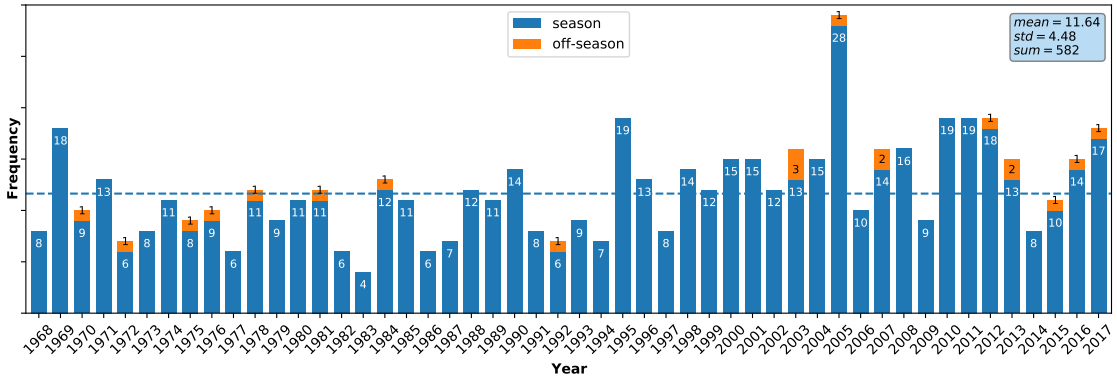

Figure 2.1: Frequency statistics of named TCs per year between 1968 and 2017, stratified into season (blue) and off-season (orange). The inset in the top-right corner presents the mean, the standard deviation, and the total number of all seasons in this 50-year period.

ambient flow with a high degree of self-sustainment, a TC not just interacts with ocean and land, but also with tropical waves (e.g., Frank and Roundy, 2006), other TCs (Fujiwhara, 1923, 1931), the stratosphere (e.g., Gray, 1984), or the mid-latitude flow (e.g., Evans et al., 2017; Keller et al., 2019; Davis and Bosart, 2003, 2004). A more detailed literature review of environmental factors influencing TC occurrence is provided in Chapter 5.

In the North Atlantic, TCs are officially named by the US National Hurricane Center (NHC) if they reach at least tropical storm strength (34 kt) during their lifetime. Since these systems largely occur in this ocean in boreal summer and autumn, an official period for the hurricane season was established in 1965, running from 1 June to 30 November. 11.64 named TCs occur during an average season, as shown by the frequency statistics of named TCs for 1968 to 2017 in Figure 2.1, whereas there are only 0.40 in the off-season (value not shown). The large standard deviation of 4.48 named TCs indicates that North Atlantic TC occurrence is influenced by atmospheric modes of interannual variability, such as ENSO. From the average number of named TCs, 6.28 and 2.50 TCs reach hurricane and major hurricane stage, respectively (values not shown). With 81 % of all named TCs occurring in the months August to October, and a peak in early September, the intraseasonal variability follows a marked seasonal cycle. This seasonal cycle results from varying oceanic and atmospheric conditions in the so-called Main Development Region (MDR), which spans from the Caribbean Sea to the eastern tropical Atlantic (80°W–20°W and 10°N–20°N). As most North Atlantic TCs originate from initial westward moving vortices, usually embedded in African Easterly Waves (AEWs), they enter this region where – under favourable conditions – a supplying air-sea interaction develops, convection becomes organized, and hence intensity increases. Apart from this classical AEW-type of North Atlantic TC genesis, McTaggart-Cowan et al. (2008) identify five

additional development pathways, based on two metrics assessing upper- and lower-level baroclinicity in a climatological study. In the analysed 1948–2004 period, the low-level-baroclinic (AEW-type) and non-baroclinic categories combine 53 % of all TCs, and are characterised by a genesis location in the MDR. The remaining 47 % primarily develop north of the MDR, and are typically influenced by strong quasigeostrophic forcing for ascent in the upper-levels. This large fraction of baroclinic influences during TC genesis leads to a prominent position of the North Atlantic compared to other basins (McTaggart-Cowan et al., 2013), and thus certainly involves more influencing factors relevant to TC forecasting.

2.1.2 Integrated tropical cyclone activity measures

Although steady progress has been made in the development of TC forecasting models over the past decades, predictions of individual TCs remain challenging. Track errors in offical NHC forecasts have decreased to such an extent that the question arose whether the inherent limit of predictability has mostly been reached (Landsea and Cangialosi, 2018). In contrast, because improvements in TC intensity forecasts are much more difficult to achieve, the US National Oceanic and Atmospheric Administration (NOAA) hast established long-term programs, such as the Hurricane Forecast Improvement Project (HFIP; Gall et al., 2013) or the Hurricane Intensity Forecasting Experiment (IFEX; Rogers et al., 2006, 2013), to unify resources and to tackle this challenge in a coordinated manner. Despite all efforts, predictive skill for individual TCs reaches its limits after a few days. To overcome this limitation and to reach longer lead times, various integrated metrics for TC activity have been developed by using integral approaches.

Gray et al. (1992) defined the Hurricane Destruction Potential (HDP) index, which sums up six-hourly instances of maximum sustained 10-m wind speed squared during the lifetime of a hurricane. By lowering the intensity threshold to tropical storm strength, Bell et al. (2000) modified the HDP index to define the Accumulated Cyclone Energy (ACE) index, given as

$$ACE = 10^{-4} \sum V_{max}^2, \tag{2.1}$$

where V_{max} is the maximum sustained 10-m wind speed. Due to the squaring of V_{max}, the calculated index is proportional to storm kinetic energy. By convention, the aggregated sum is multiplied by 10^{-4} to obtain smaller values, and is usually presented without units (in $10^4 kt^2$). The ACE index can be calculated for a single storm but also for all storms occurring in a given season and ocean basin. The seasonal integration of ACE yields a measure that is predicted by several institutions for seasonal outlooks of TC activity, collectively presented under `https://seasonalhurricanepredictions.bsc.es`. Using

seasonal integration, Drews (2007) decompose ACE into multiplicative factors for number of storms per season, seasonally averaged duration, and seasonally averaged intensity, respectively. Applied to the North Atlantic 1851-2006 period, he finds that interannual changes in TC activity are dominated by the change in number of storms, and not so much by changes in the other two factors.

Replacing the exponent of V_{max} in Eq. 2.1 by N yields a generalized formulation that can be seen as a template for the definition of other integrated TC activity indices in the literature. For $N = 3$, the obtained measure is referred to as Power Dissipation Index (PDI; Emanuel, 2005). This formulation goes back to the approach of deriving TC power dissipation by spatial integration over the dissipation rate (Emanuel, 1999), which is proportional to V_{max}^3 and mostly confined to the surface layer (Bister and Emanuel, 1998). As V_{max}^3 is also approximately proportional to monetary loss (Emanuel, 2005), the PDI is a measure particularly useful for the insurance sector. For $N = 0$, the generalized version of Eq. 2.1 reduces to a plain summation of storm days, a measure for the duration of TC activity.

2.2 Dynamical modelling

2.2.1 Basic concept

Dynamical models attempt to describe the time-dependent behaviour of a system trough simplified representation. The underlying processes, generating observable variables, are usually governed by differential equations. When modelling the atmosphere, the following fundamental set of governing equations is borrowed from the field of fluid dynamics (e.g., Doms and Baldauf, 2021). The first equation, referred to as Navier-Stokes equation, is given by

$$\rho \frac{d\vec{v}}{dt} = -\rho \vec{\nabla} \phi_G - \vec{\nabla} p - \vec{\nabla} \cdot \Psi \tag{2.2}$$

where ρ is the air density, $\vec{v}$ is the wind vector, t is time, ϕ_G is the gravitational potential, p is the air pressure, and Ψ is the stress tensor due to viscosity. Equation 2.2 results from Newton's second and third axioms, namely that a force acting on a body can be written as the time derivative of its momentum, which equals zero when summed over all forces, thus expressing conservation of momentum. The forces considered in meteorological models are due to gravity, pressure, and friction.

Another type of equation in this set is the so-called continuity equation. It generally states that the change of an intensive (i.e., non-mass proportional) quantity a within a given volume is determined by the flux of a ($\vec{I_a}$) through the volume's surface and by the

generation or loss of a (σ_a) in the interior.

$$\rho \frac{da}{dt} = -\nabla \cdot \vec{I_a} + \sigma_a \tag{2.3}$$

For $a = 1$, the second term on the right-hand side vanishes since mass is conserved, and Eq. 2.3 can be written as

$$\frac{d\rho}{dt} = -\rho \nabla \cdot \vec{v}. \tag{2.4}$$

The atmosphere is as a multicomponent system, and usually modelled consisting of dry air (d) and water in form of vapour (v), liquid (l), and ice (i). Each of these components $c = \{d, v, l, i\}$ can be expressed as a mass fraction $x_c = \rho_c/\rho$, resulting in intensive quantities, for which Eq. 2.3 yields four additional equations.

$$\rho \frac{dx_c}{dt} = -\nabla \cdot \vec{I_{x_c}} + \sigma_{x_c} \tag{2.5}$$

According to the first law of thermodynamics, formulated for a closed, homogeneous system, energy is conserved. This constitutes the foundation for the last prognostic differential equation used to model atmospheric flow. Using the formulation based on specific enthalpy h, expanding the $h = h(T, p, x_c)$, where T is the air temperature, and finally substituting dp/dt by the pressure tendency equation, one finally yields the following heat equation

$$\rho c_v \frac{dT}{dt} = -p \nabla \cdot \vec{v} + Q_h + Q_m, \tag{2.6}$$

where c_v is the specific heat at constant volume. Q_h and Q_m are production terms associated with diabatic heating and moist processes, respectively, which consider influences from heat and diffusion fluxes, phase transitions and radiation fluxes.

To close the set of equations for $\vec{v}, \rho, x_v, x_l, x_i, T$ and, p, the equation of state is used with the assumption that dry air and water vapour are ideal gases whereas liquid water and ice are incompressible. This provides a diagnostic equation for p of the following form

$$p = \rho \left(R_d x_d + R_v x_v \right) T, \tag{2.7}$$

where R_d and R_v are the specific gas constants for dry air and water vapour, respectively.

As the earth rotates at an angular speed of $7.292 \times 10^{-5}\,\mathrm{rad/s}$, so does the atmosphere. An earth-relative formulation of the equations of motion therefore necessitates a transformation to a non-inertial, uniformly rotating frame of reference. As a consequence, in addition to the term of earth-relative acceleration, two additional terms appear in the momentum equation, the Coriolis force and the centrifugal force. These are fictitious forces, which only appear due to the chosen frame of reference, and thus do not

have independent influence on a body of mass. As described in Section 2.1.1, these two forces are part of the gradient wind balance that determines the primary circulation of a TC. In other words, TCs would not occur on a non-rotating earth. In addition to the earth-relative formulation, the set of equations is transformed to a spherical coordinate system to allow for a more intuitive definition and convenient use of the model.

Because analytical solutions do not exist, numerical integration methods are required to solve the set of equations. To achieve this, the original continuous formulation of the partial differential equations needs an appropriate discretization in time and space, in a way that solutions are stable. In the course of discretization, most NWP models introduce a certain type of grid for the horizontal dimensions, and a terrain-following coordinate in the vertical. Alternatively, some models solve the set of equations in spectral space. Depending on the chosen grid spacing, physical processes that cannot be resolved must be parameterized, so that the influence of those subgrid-scale processes can feed back to the predicted variables at the defined grid points. Common parameterization schemes exist for convection, radiation, clouds, precipitation, turbulence, surface layer, soil, sea-ice, and others (e.g., Doms et al., 2021).

The model domain is bounded at the bottom by the land and sea surface, while a maximum height is usually defined at the top. While limited area models rely on information at the lateral boundaries, such a dependency does not exist for global models, however, solutions must be periodically identical in a circumglobal sense (e.g., Schättler and Blahak, 2021). The integration of the differential equations further necessitates initial conditions for all variables at each model gridpoint. This part of the modelling process is computationally very expensive, since observations are inhomogeneously spread in time and space, and sometimes even do not measure the predicted variables directly. Data assimilation techniques are therefore required to determine the atmospheric state at the time of forecast initialization (e.g., Schraff and Hess, 2021). The underlying conceptual approach in these techniques is that the initial state of the atmosphere is estimated through interpolation between the observations and a first guess of the model state, which is usually obtained from previous forecasts.

Due to errors associated with observations, parameterization schemes for unresolved processes, numerical methods, and other model-related sources, forecast errors typically increase with lead time. Even in case of a perfect deterministic model and any small errors in initial conditions, the chaotic nature of the atmosphere would inevitably lead to forecast errors that limit predictability. Since the very beginning of the operational use of NWP models, single deterministic forecasts were issued per initialization time, before a new concept of forecast production began to prevail about three decades ago. Nowadays, an ensemble of forecasts is usually computed, with each member starting from slightly, but reasonably perturbed initial conditions (Leutbecher and Palmer, 2008).

With the assumption that all members are statistically consistent, i.e. they behave like independent random realisations of the variables to be predicted, the sample distribution allows to infer statistical properties of the underlying predicted distribution. Based on the dispersion of the ensemble forecast, usually measured be standard deviation, forecast uncertainty can be quantified.

2.2.2 Tropical cyclone forecasting

Until recently, grid spacing in global NWP models was too coarse to resolve and realistically predict TCs. Based on a sensitivity study, in which forecast experiments of Hurricane Ivan (2004) were run for a range of grid spacing values (8 to 1 km), Gentry and Lackmann (2010) showed that distinct eyewall segments with localized updrafts and more pronounced spiral bands were resolved beginning at about 4-km grid spacing. When going to even smaller grid spacing, finer structures related to convective processes in the eyewall were found to be better resolved, leading to their suggestion that a grid spacing of 3 km or less is required for operational purposes of TC forecasting.

Since such high resolution forecasts are not produced operationally in global NWP models to date, TC-following nests with grid spacing finer than for the outer domain have been developed to allow for high resolution modelling in the TC environment. An example for this approach, which became operational in 2007, is the Hurricane Weather Research and Forecast (HWRF) model developed by the National Centers for Environmental Prediction. A key advantage of this configuration is that airborne in-situ observations (e.g., dropsondes, flight-level measurements, tail-Doppler radar) from reconnaissance flights can be assimilated by means of tailored data assimilation techniques. This allows to replace the vortex of the global model with a TC initialised with a more realistic structure and intensity. Further improvements through such a nesting approach can be achieved by ocean initialisation and coupling, and refinement of parameterization schemes (e.g., Doyle et al., 2014).

Due to steady improvements in model development and recent advances in computing, the current generation of global models is able to treat convection explicitly (i.e., grid spacing is about 5 km or less), and hence to resolve TCs (Judt et al., 2021). This ability raises the prospect of a new stage in short- to medium-range TC modelling, particularly with respect to intensity prediction, although the long-standing problem of rapid intensification may in some cases require exceeding the 1-km mark for grid spacing (Fox and Judt, 2018).

2.3 Statistical modelling

2.3.1 Basic concept, terminology, and assumptions

As an alternative to dynamical models run in ensemble mode for an assessment of forecast uncertainty, statistical (or stochastic) models can be used to predict the outcome of random variables in the form of probability distributions. The modelling of the data-generation process is based on link functions that mathematically describe the relationship between the random variable(s) to be predicted and some non-random variables. This way, statistical models can be developed even if knowledge about the underlying physical processes is lacking or too limited. In case of quantitative variables, the variable type can be either discrete or continuous. Depending on the field of application, independent variables are also referred to as predictors, features, explanatory variables, or covariates, whereas dependent variables are also called predictands, target variables, or explained variables. In the context of forecasting, we hereafter mainly use 'predictors' and 'target variable', while instances of the latter are also referred to as 'observations'.

The formulation of the link function requires choices and assumptions to be made to model the relationship between predictor and target variables. A model is referred to as 'univariate' if only one predictor is used, whereas a 'mulitvariate' model depends on more than one predictor. An assumption often made, but not always justified, is that the variance of the target variable is independent of the value of one or more predictors. In this case the model is called 'homoscedastic', and 'heteroscedastic' otherwise. Another important choice is whether a parametric or nonparametric model should be used. Contrary to what the name suggests, nonparametric models also have parameters, but these are determined implicitly from data and are not fixed a priori. By contrast, a parametric model prescribes a finite number of parameters that shapes a theoretical distribution to best fit the distribution of the target variable. However, in case the underlying data-generating process is not really understood, there is no way to readily determine which theoretical distribution is most appropriate. Although a number of candidate distributions can often be narrowed down from the wide range of theoretical distributions that exists in the literature (Krishnamoorthy, 2006) through preliminary considerations, the choice remains subjective and can only be conclusively evaluated through validation of the model's predictive performance. A good and commonly practiced strategy for selection is to test heuristic approaches and keep the model as simple as possible, i.e., to minimize the number of model parameters.

Once choices and assumptions are made, model parameters need to be determined by fitting the link function to a dataset, which consists of a finite number of instances for the predictor and target variables. Because the statistical population of any atmospheric

variable is generally unknown, samples are drawn to represent the underlying population in statistical analysis and modelling. Technically, these sample datasets are constructed from observations, i.e., from past realisations of the variables. Using these data, the set of model parameters can then be estimated by a solver through minimization of a loss function. From that point on, the model is determined and it can be used to predict the target variable based on values provided for the predictors.

Apart from being deployed as stand-alone forecast models, statistical models are also used to correct for systematic errors in predictions of dynamical models, which is referred to as 'statistical post-processing'. Nowadays, there is a wide range of techniques for this purpose that continues to grow due to new areas of research and the big data-driven demand of many applications (e.g., Vannitsem et al., 2021). Due to the diversity of target variables and the form in which they are provided (i.e., point forecasts, predictive distribution, etc.), the field of meteorological forecasting contributes significantly to the improvement of existing and the development of new statistical post-processing methods.

2.3.2 Tropical cyclone forecasting

As discussed in 2.1.2 and 2.3.1, predictions for TCs are challenging in general, but especially with regard to their intensity. Therefore, statistical models have been developed over the past decades to compensate for the deficiencies of dynamical models in TC intensity forecasting. An example that was made operational in 1990 is the Statistical Hurricane Intensity Prediction Scheme (SHIPS; DeMaria and Kaplan, 1994; DeMaria et al., 2005) model, which predicts maximum sustained winds, at that time, using climatological, atmospheric, and oceanic information along with persistence to generate predictors. Other applications for individual TCs that use statistical models have been developed, e.g., for predicting the probability of rapid intensification (Kaplan et al., 2010), for TC tracks (e.g., Hall and Jewson, 2007), and changes in TC structure, as well as for the eyewall replacement cycle (Kossin and Sitkowski, 2012).

Shifting the perspective from individual TCs to integrated TC activity of entire seasons, statistical models have long been the only means for producing seasonal outlooks (Klotzbach et al., 2017). The basis of their development is the identification of statistical correlations between TC activity and different modes of atmospheric and oceanic variability. Once significant correlations are found, predictors can be created to forecast various measures of TC activity, with predictions often categorized into above-normal, normal, or below-normal.

For subseasonal leadtimes, inspired by the example of numerous seasonal models, Leroy and Wheeler (2008) developed a logistic regression model 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 forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) model, 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 bias-corrected version of the ECMWF forecasts, as well as against the average of predictions from the statistical and the bias-corrected ECMWF models, which further improved skill. For the North Atlantic, Henderson 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 Henderson 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.

2.4 The subseasonal predictability gap

As forecast skill of dynamical models is limited to a couple of days, and statistical models have demonstrated value especially for seasonal predictions, the development of these two distinct approaches proceeded in parallel for decades. This coexistence is due to the subseasonal predictability gap (Vitart et al., 2012; Robertson et al., 2020), which has raised broad attention and efforts to bridge only in recent years. While dynamical models obtain their predictive power from the initial conditions of the atmosphere, seasonal outlooks exploit information from boundary conditions, primarily from sea surface temperatures (SSTs). The gap in predictability between the medium-range and the seasonal regimes is a long-standing problem in weather forecasting and also affects TC predictions.

The recent progress in high-performance computing not only enables to run forecasts at high resolution, but also to extend the forecast range to several months. To provide a common database of such extended NWP forecasts that are produced by different operational centers, and to create a platform that facilitates the coordination of research efforts for this time horizon, the Subseasonal-to-Seasonal (S2S; Vitart et al., 2017) project was initiated. The S2S database consists of extended forecasts from 11 forecast centers with lead times up to 62 days. The models differ in their configuration, e.g. resolution (about $0.25°$–$2°$), ensemble size (4 to 51 members), initialisation frequency (daily to weekly), and whether they are coupled to an ocean and/or sea-ice model. In addition to operational forecasts, smaller ensembles of reforecasts for up to 32 years are also provided,

allowing for forecast evaluation over longer time periods. Although the S2S project is motivated by a range of scientific questions, two key goals are to evaluate forecast skill and to identify potential sources of predictability for the subseasonal timescale.

An overview of potential sources of subseasonal predictability is given by Vitart et al. (2012), who list the MJO, soil moisture, snow cover, stratosphere-troposphere interaction, and ocean conditions. For the latter source, influence can either be local or remote via atmospheric teleconnections, as is the case for the prominent example of ENSO. With the growing understanding of intraseasonal to interannual atmospheric and oceanic modes of variability, the MJO and ENSO were found to have modulating influence on subseasonal TC activity (Camargo et al., 2019). The MJO and ENSO and their impact on TC activity is discussed in more detail in Chapter 5. The influence of the MJO has not only been identified in observational data but also shown in subseasonal NWP forecasts for TC activity in many oceans (e.g., Vitart, 2009; Belanger et al., 2010; Camp et al., 2018). Since the MJO can be predicted up to 5 weeks in advance (Kim et al., 2018), expectations are high that subseasonal TC predictions benefit from this skill.

In rare cases, forecasts for genesis and intensity of individual TCs proved to be correct up to about one week ahead. For the Severe Cyclone Nargis in the North Indian Ocean in 2008, for example, it was possible to predict its formation 8 days and its hurricane intensity 6 days in advance (Belanger et al., 2012). However, despite the ability of dynamical models to skillfully forecast the MJO up to subseasonal leadtimes, predictive skill of the occurrence of individual TCs largely vanishes after forecast week 1 compared to climatological predictions (Lee et al., 2018). Evaluating subseasonal reforecasts from six of the eleven S2S models in terms of weekly TC occurrence, Lee et al. (2018) find that the ECMWF model performs best overall. In the North Atlantic, basin-wide probabilistic forecasts from this model slightly exceed predictions based on a monthly varying climatology up to week five. By using each reforecast ensemble member once as an observation for verification of the remaining ensemble, and then averaging results, Lee et al. (2018) aim to assess potential predictability (Buizza, 1997), a measure that gives an indication of room for improvement in dynamical modelling. Based on this method, the actual predictive skill of the ECMWF model for North Atlantic weekly TC occurrence has almost reached its limits. This suggests that no major improvements can be expected for subseasonal TC occurrence prediction with the current generation of NWP models. In a follow-up study, Lee et al. (2020) evaluate TC occurrence within gridded $20°\times15°$ regional boxes instead of the previous basin-wide approach. Due to these spatially more confined evaluation regions, the ECMWF predictive skill already drops below the performance of climatological predictions after week one, which confirms the challenge to predict TCs at subseasonal leadtimes.

2.5 Statistical-dynamical modelling

The information contained in the initial conditions, which is the fundamental source of predictability in dynamical modelling, becomes blurred as lead time increases. An approach that is considered promising for extending predictive skill beyond the limits of dynamical modelling is statistical-dynamical (or hybrid) modelling. The conceptual idea behind this is to use the predictive skill contained in dynamical forecasts to generate predictors that are then used in a statistical model. The way of how this is technically realised differs strongly on the application, and ranges from man-made specifically developed predictors to machine-learned hidden representations of the underlying data. Although the hybrid approach has been used for TC forecasting in several applications, the vast majority of models target either medium-range or seasonal lead times, so that the potential value of this approach for subseasonal TC predictions has remained virtually unexplored.

An example of hybrid forecasting on the medium-range is given by the SHIPS model (see Section 2.3.2), which predicts TC intensity up to 5 days ahead. Originally developed as a purely statistical model, additional NWP-based predictors were incorporated as of 1997 (CIRA RAMMB, 2022), changing it from a purely statistical to a statistical-dynamical model. On the seasonal range, statistical-dynamical models can nowadays also be developed for outlooks of TC activity due to the availability of NWP forecasts integrated up to several months, thus offering an alternative approach to the purely statistical models so far. While the seasonal hybrid models differ in the dynamical predictors used and how the statistical model is defined, most of them have in common that predictions are basin-wide and seasonally integrated, e.g. for TC frequency and landfall (Murakami et al., 2016; Zhang et al., 2017b), or ACE (Villarini et al., 2019; Klotzbach et al., 2020). The ability of the hybrid approach to improve seasonal forecasts of TC counts for lead times up to 7 months has been demonstrated by comparison to direct predictions from the underlying dynamical model (Murakami et al., 2016; Zhang et al., 2017b). For seasonal North Atlantic ACE, the statistical-dynamical model developed by Klotzbach et al. (2020) showed skill when combined with predictions from their pre-existing statistical model. A detailed overview of the various seasonal forecast models and differences in their approaches, target variables, and ocean basins is presented in Klotzbach et al. (2019). Finally, it is worth to note that hybrid modelling has also been successfully applied in the context of multi-annual predictions of ACE (Caron et al., 2015).

3. Research questions

The previous chapter provided an overview of different types of modelling approaches and reviewed how they are used in several applications for TC forecasting. Since with the current generation of NWP models only minor improvements are expected for subseasonal TC activity forecasts in the North Atlantic (Lee et al., 2018), any improvements in predictive performance through a statistical-dynamical approach can be more clearly attributed to the value of adding a statistical component. Therefore, the development of a hybrid model and its systematic comparison against other approaches will be confined to selected subregions of the North Atlantic basin. While different measures are defined and used in the literature to describe TC activity, the focus in this thesis is on two widely considered target variables that cover key forecasting aspects. The first target variable is concerned with the occurrence of TCs. Building on this foundation, the integrated measure of TC activity, ACE, is addressed as the second target variable, which incorporates the additional aspect of TC intensity. The overarching research aims outlined in Chapter 1 guide the design of this study and the formulation of specific research questions motivated in the remainder of this chapter. The first part is dedicated to the identification of subseasonally relevant predictors to be generated from NWP forecasts, according to the hybrid approach. Using these predictors, models for TC occurrence and ACE are developed and validated in the second and third part of this study, respectively.

The introduction section of scientific articles addressing TC formation often begins with a reference to some early studies that set out to deduce necessary environmental conditions (e.g., Palmen, 1948; Gray, 1968). Although there is a certain consensus in the TC research community regarding the list of such conditions, the necessity of some of these conditions is still under debate and alternative factors are occasionally proposed. To date, such studies are mostly based on observational or (re)analysis data, which is why these genesis factors represent a direct physical influence, both temporally and spatially. With the aim of predicting at subseasonal leadtimes, the question arises whether (all) these commonly accepted environmental factors can be practically used as predictors when generating from dynamical forecasts of the underlying variables. For remote influences, such as from ENSO, it is also unclear which forecast time of the dynamical

component is best to choose for predictor generation. Because the hybrid approach is fundamentally based on the assumption that these influencing factors are predicted sufficiently well, this needs to be verified first. Although some studies suggest that subseasonal TC forecasts would benefit from predictions of atmospheric modes of variability, e.g. from tropical waves or the MJO (Janiga et al., 2018; Frank and Roundy, 2006), no analysis has yet provided an overview of the subseasonal relevance of a wide range of factors to TC prediction. Chapter 5 therefore conducts such a comparison study to verify the above assumption. This part of the study will present a literature review and physical motivation of relevant predictors to provide an extensive predictor pool from which the statistical model component can then select an optimal subset in Chapters 6 and 7. In addition to the predictors used in previous studies (Leroy and Wheeler, 2008; Vitart et al., 2010; Henderson and Maloney, 2013), further potentially relevant predictors, related to the genesis potential index (GPI), tropical waves, and extratropical dynamics, will be included. Using a subseasonal lead time, Chapter 5 finally analyzes patterns of correlation with TC occurrence to discuss in which subregions predictors are most useful. The subseasonal relevance of predictors is therefore addressed by the following research questions:

RQ 1a What factors influencing TC occurrence are known in the literature, and are they likewise relevant when corresponding predictors are generated from subseasonal NWP forecasts? (Chapter 5)

RQ 1b What are the key predictors selected by the hybrid model at each forecast week when predicting TC occurrence and ACE, respectively? (Chapters 6+7)

RQ 1c How much does each predictor group contribute to the predictive skill of the hybrid model? (Chapter 6)

To convey information on forecast uncertainty, all models used in this thesis produce probabilistic forecasts. While a single probability is output for TC occurrence, the full predictive distribution is modelled for ACE. For both target variables, weekly forecasts are issued for the first five consecutive weeks. Instead of basin-wide predictions, as is usually done in seasonal forecasts, a gridded framework is deployed to gain insight into subregional differences. Beyond the development of the hybrid approach, an important contribution of this dissertation is to compare a variety of different model types in terms of predictive skill in a systematic way. For TC occurrence models (Fig. 3.1), the climatological seasonal cycle of TC occurrence probability (gray) constitutes the first benchmark, which is further optimized by a suitable smoothing method, resulting in a second bench-

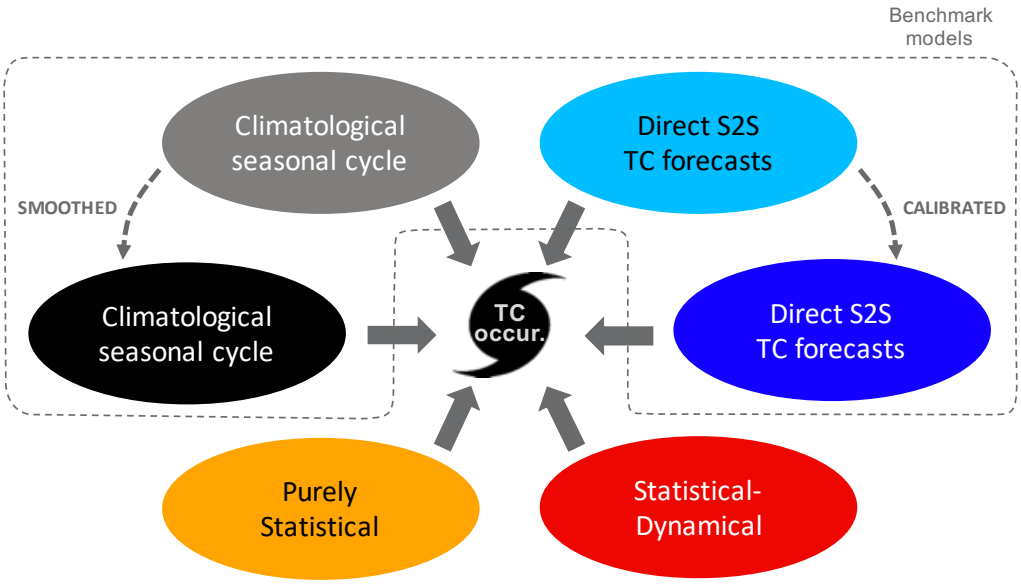

Figure 3.1: Schematic overview of probabilistic models forecasting TC occurrence. Benchmark models are enclosed by the gray dashed contour.

mark (black). A fundamentally different approach is taken by deriving TC occurrence from S2S ensemble forecasts (light blue). As NWP predictions can suffer from biases, statistical post-processing is used to create a calibrated model version of it (dark blue). Following the idea of Leroy and Wheeler (2008), namely to generate relevant predictors from past data, the performance of this purely statistical approach (yellow) is compared to the other benchmark models first. Using the same predictor types but generated from NWP predictions, the utility of a statistical-dynamical approach (red) is examined in a final step. Models developed and compared for ACE are analogous to their TC occurrence counterparts in the underlying approach (Fig. 3.2), but they differ in that the predictive distribution modelled requires an adapted formulation. In addition, a trivial model predicting TC non-occurrence with certainty is included as another benchmark (green). Detailed descriptions of how each individual model is set up are presented in Chapters 6 and 7, respectively.

As reviewed in the Chapter 2, previous studies have typically developed and validated individual of these distinct modeling approaches for subseasonal forecasting. But only rarely have they been directly compared with each other. A major problem is that uniform standards regarding formulation of target variables, lead times, and verification metrics do not exist, making it virtually impossible to compare models from different studies and institutional groups (Camargo et al., 2019). Vitart et al. (2010) evaluate forecasts of a climatological and a dynamical model with the purely statistical model of Leroy and Wheeler (2008) for the Southern Hemisphere, but they neither optimize the climatology used as the basis for the statistical model nor consider a hybrid approach. The present study is the

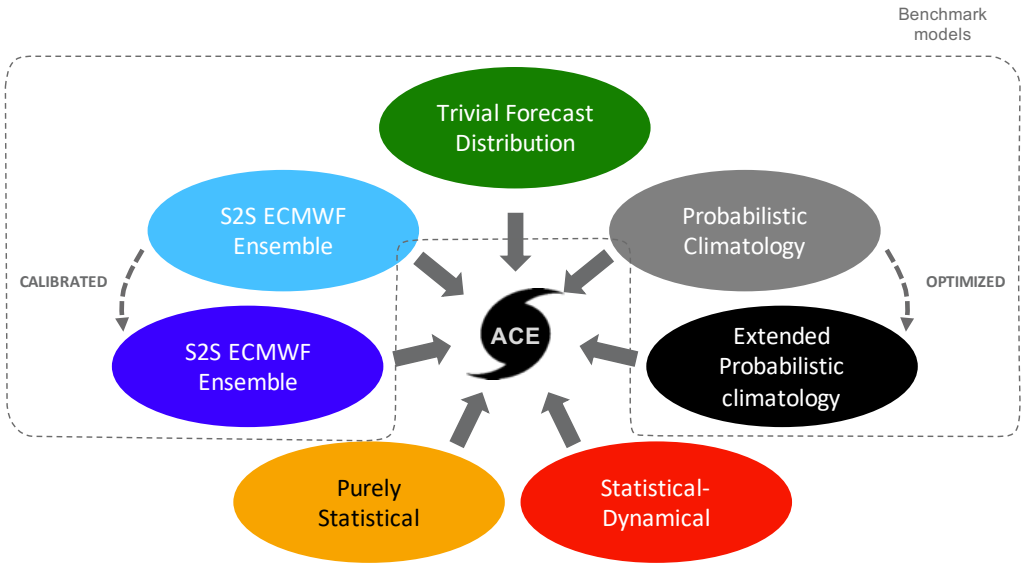

Figure 3.2: Same as Fig. 3.1, but for probabilistic models forecasting the predictive distribution of ACE.

first to develop and systematically validate subregional forecasts for North Atlantic TC activity out to week five of i) a statistical–dynamical approach, ii) a purely statistical approach (as in Henderson and Maloney, 2013), iii) different climatological models, as well as iv) (un)calibrated dynamical models at once. The probabilistic forecasts for the two target variables require different verification approaches, which are described in detail in Section 4.4. The single probability forecasted for TC occurrence poses a binary problem in terms of forecast evaluation, whereas the predictive distribution of ACE requires more complex approaches. Model validation in this study is conducted in two subregions of the North Atlantic: the Gulf of Mexico (represented by 20°N–30°N, 100°W–80°W), and the central MDR (represented by 10°N–20°N, 60°W–40°W). The two subregions differ especially in that TC activity in the central MDR is subject to a stronger seasonal cycle. While TCs in the Gulf of Mexico are frequently exposed to strong upper-tropospheric baroclinicity associated with the mid-latitude flow, TCs occurring in the central MDR often interact with AEWs, thereby experiencing low-level baroclinicity (McTaggart-Cowan et al., 2008, see their Fig. 12). The validation performed evaluates various forecasting aspects to gain a comprehensive assessment. Forecast models are examined in terms of reliability as well as of potential and actual predictive skill. Furthermore, decomposition of evaluation scores and analysis of model's economic value provide deeper insight into forecast quality and allow to identify model weaknesses and limitations. For the two target variables and two subregions, the following research questions addressed are:

RQ 2a How well can each model discriminate between TC occurrence and non-occurrence[1], and are their forecasts calibrated?

RQ 2b How does the dynamical model perform over the five forecast weeks considered, and can statistical post-processing help improve the predictive performance?

RQ 2c In comparison, at which forecast week does the climatological model become more skillful, and is it worth optimizing its representation?

RQ 2d Can the purely statistical modelling approach, using past data to generate predictors, exceed the skill of the climatological model at subseasonal lead times?

RQ 2e Does the statistical-dynamical approach, i.e. generating the same predictors from NWP forecasts, actually yield the putative subseasonal improvements?

RQ 2f Which model provides the highest value for economic decision making at each forecast week?[1]

[1]Note that this question is only addressed in the context of TC occurrence modeling (Chapter 6), as the underlying validation concepts are not applicable to non-binary target variables.

4. Data and methods

4.1 Data and derived products

4.1.1 Target variables

All models developed and evaluated in this thesis forecast one of two variables considered (hereafter referred to as target variables), namely either TC occurrence or ACE. Both variables are derived from the International Best Track Archive for Climate Stewardship (IBTrACS; Knapp et al., 2010, 2018) dataset version 4. Since the IBTrACS dataset is compiled from many sources, tracks may vary between agencies and TC positions are therefore labelled as either 'main' or 'spurs'. The latter are omitted to avoid biases caused by multiple representations of TC tracks. A track in this dataset can cover a range of phases during a cyclone's lifetime, and because this study focuses on tropical cyclones, only track portions are used to generate the target variables, for which the cyclone fulfills two requirements. Firstly, the cyclone is required to be tropical in nature, and secondly, it's intensity has to reach at least tropical storm strength ($\geq 34\,\mathrm{kt}$). 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 dataset, the TC tracks in the S2S model (cf. Section 4.1.3). Figure 4.1a 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. The majority of TC positions is found over the oceanic regions between $10°$N–$50°$N, and $100°$W–$20°$W, with highest densities occurring in the western part of the North Atlantic.

An evaluation area of some size is needed to create a regularly gridded dataset for each target variable from the set of irregularly distributed TC positions. The choice of a coarser spatio-temporal evaluation is further reasonable to account for the lower predictability on subseasonal timescales. The target variables at a given gridpoint are therefore defined by evaluating the TC positions in a circular area of radius $7.5°$ centered on the that gridpoint (Fig. 4.1b), and within each forecast week. Different radii of the evaluation circle were tested, but $7.5°$ was chosen as a good compromise between including more TC positions

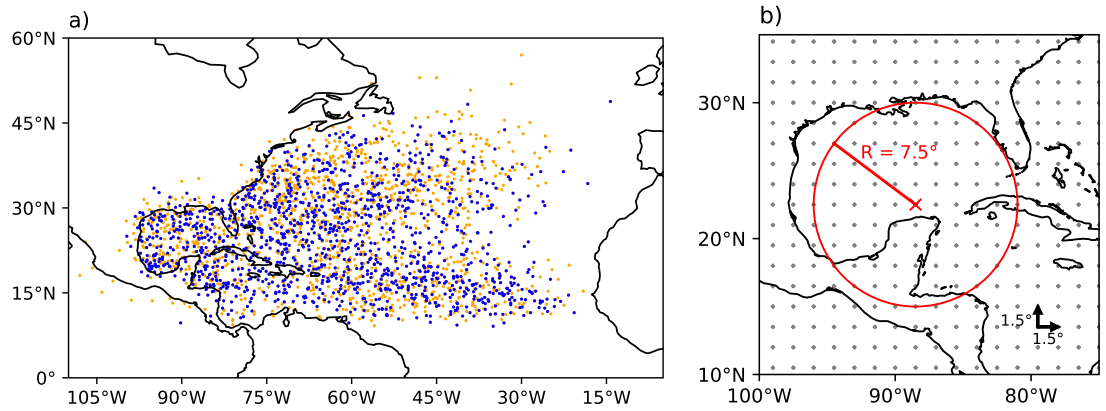

Figure 4.1: (a) 1968-1997 (orange dots) and 1998-2017 (blue dots) IBTrACS TC positions at 0000 UTC during the North Atlantic hurricane season (June-November) for intensities of at least tropical storm strength. Reprinted from Maier-Gerber et al. (2021). © 2021, American Meteorological Society. Used with permission. (b) Illustration of the $7.5°$-radial TC evaluation, used to generate the target variables, at an example grid point in the Gulf of Mexico. This technique is applied at every grid point of the $1.5° \times 1.5°$ regular grid (gray dots).

(i.e., using a larger radius) to obtain less extreme target variables, and minimizing spatial uncertainty about the exact TC position (i.e., using a smaller radius). This evaluation is done at every gridpoint of a $1.5° \times 1.5°$ regular grid, which is used to compare with forecasts from the coarsest spatially resolved dataset, the TC tracks in the S2S model (cf. Section 4.1.3). Hence, each evaluation circle overlaps with the nearest gridpoints by about 87% of its area, resulting in a high local consistency for the target variables between neighbouring gridpoints.

The first target variable, TC occurrence, applies this evaluation strategy and assigns a 1 (i.e., TC occurrence) to a particular gridpoint, if at least one TC occurs within the corresponding circular area over the forecast week considered. Otherwise, the gridpoint is assigned a 0 (i.e., TC non-occurrence). Based on this binary target variable, Figure 4.2a presents a map of relative frequencies of TC occurrence. The overall pattern naturally resembles the density of TC positions that can be visually assessed in Fig. 4.1a. A pronounced maximum of high relative frequencies above 15% occurs in the Gulf of Mexico, with values slowly decreasing in two bands towards Europe and southwest Africa, respectively. Given the extreme nature of TCs, the ratio of TC occurrence and TC non-occurrence instances should not be too imbalanced for statistical models to be properly trained. Except for northeastern part, North Atlantic TC occurrence appears with a fraction of more than 1% during the hurricane season, as indicated by the red contour in Fig. 4.1a. For the Gulf of Mexico and central MDR subregions, which are defined for model validation (cf. Section 4.3), the gridpoint-averaged relative frequency of TC occurrence is even 13.3% and 11.8%, respectively. Thus, this target variable is not so imbalanced

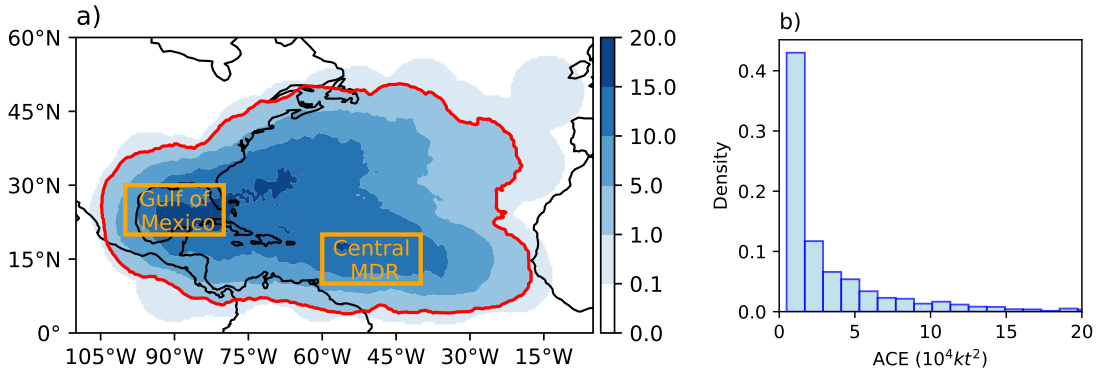

Figure 4.2: (a) Relative frequency of TC occurrence (%) calculated for the 1998-2017 hurricane seasons (see text for details on how TC occurrence is defined). Note that interval boundaries are not equidistant. The red contour encloses the area where TCs occur at a rate of more than 1%. Orange boxes enclose the subregions used for model validation. Reprinted from Maier-Gerber et al. (2021). © 2021, American Meteorological Society. Used with permission. (b) Histogram of ACE > 0 (in $10^4 kt^2$) for the 1998-2017 hurricane seasons and gridpoints pooled within the central Main Development Region subregion.

that statistical models would have too few instances of TC occurrence to learn from.

The second target variable generated, ACE, differs from its original definition in three ways (cf. 2.1.2). The first deviation is that instead of the usual basin-wide aggregation, the above described gridded framework for TC evaluation is applied to allow for model development, validation and comparison in different North Atlantic subregions. Unlike the binary evaluation perspective of TC occurrence, ACE sums the maximum wind speed squared at all TC positions of all TCs occurring within the spatio-temporal evaluation domain, divided by the conventional factor of 10^4. The application of the gridded framework is further corroborated by the fact that days with two or more TCs in the North Atlantic basin only occurred on $4\,\%$ of all days during the 1968-2017 hurricane seasons. A basin-wide evaluation is thus too coarse on most days anyway. A second deviation from the original approach of calculating ACE is that only daily (0000 UTC) values are aggregated, instead of the conventional six-hourly values. This is necessary to consistently compare with ACE calculated from TC tracks in the S2S model. To correct for the larger aggregation interval and to approximate typical values, the daily aggregated ACE is multiplied by a factor of 4. While the annual mean (standard deviation) of the 1968–2017 North Atlantic 6-hourly aggregated basin-wide ACE is $99.42 \times 10^4\,\mathrm{kt}^2$ ($59.93 \times 10^4\,\mathrm{kt}^2$), it is $99.26 \times 10^4\,\mathrm{kt}^2$ ($59.97 \times 10^4\,\mathrm{kt}^2$) using the calculation approach applied here. The high agreement between the numbers indicates that the conventional ACE can be well approximated by the 0000 UTC instances. The third deviation from the original approach is that subtropical phases are not considered for the aggregation, to allow for a consistent calculation of ACE among all models in this thesis. Using the de-

scribed approximation for ACE, the annual mean (standard deviation) of the 1968–2017 aggregated basin-wide ACE for tropical stages only is $96.76 \times 10^4\,\mathrm{kt}^2$ ($60.50 \times 10^4\,\mathrm{kt}^2$). The marginal deviations from the values above suggest that the main statistical properties of the approximated ACE are very close to the original ones, so that aggregating tropical phases only is well justified.

As mentioned earlier, the central MDR gridpoints exhibit an averaged relative frequency of TC occurrence of 11.8%. For this fraction of cases, Figure 4.2b shows the distribution of ACE, with high probability densities for very low ACE values and a pronounced decline with increasing ACE. A key challenge for models in predicting this target variable hence is to forecast distributions, which strongly resemble the described shape on average. Due to the required tropical storm strength, ACE takes on 0 or values $\geq 4 \times 10^{-4} \times (34\,\mathrm{kt})^2 = 0.4624 \times 10^4\,\mathrm{kt}^2$, resulting in a small gap that must be considered when developing statistical models.

4.1.2 Predictor variables for the statistical models

The difference between the statistical-dynamical approach and the purely statistical approach lies in the underlying data, from which predictors are generated. The purely statistical model is trained on ERA5 reanalysis data (Hersbach et al., 2020), whereas predictors for the statistical-dynamical model are generated from S2S ECMWF ensemble reforecasts. For the latter, we use model version dates from the 2018 North Atlantic hurricane season, which means that the corresponding reforecasts belong to the 1998–2017 seasons. While the ERA5 dataset was produced using model version Cy41r2 of the Integrated Forecasting System (IFS) – ECMWF's atmospheric model and data assimilation system – the S2S reforecasts were based on Cy45r1. Despite some changes (e.g., in data assimilation, atmosphere–ocean coupling, and parametrization 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, running to 46 days ahead. Originally calculated with a horizontal grid spacing of $16\,\mathrm{km}$ for the first 15 days and $31\,\mathrm{km}$ 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).

$$PV = -g\left(\frac{\partial u}{\partial p}\frac{\partial \theta}{\partial y} - \frac{\partial v}{\partial p}\frac{\partial \theta}{\partial x} + \frac{\partial \theta}{\partial p}(\zeta + f)\right) \tag{4.1}$$

where g is the gravitational acceleration, u and v the horizontal wind components, p the pressure, θ the potential temperature, ζ the relative vorticity, and f the Coriolis parameter. Even though intervals between pressure levels are rather large in the S2S dataset, affecting the calculation of vertical derivatives and thus PV, they are yet thought to be sufficient for generating predictors, as this work primarily seeks to represent the integral effect of PV rather than finding an optimal representation of PV objects.

To ensure that the S2S-based predictors are not subject to biases, a mean bias correction is applied to all variables, from which predictors are directly generated. Using the S2S reforecasts of the 1998-2017 seasons, mean biases are calculated with respect to the temporally corresponding ERA5 data, 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.

4.1.3 Tropical cyclones in S2S data

Tropical cyclone tracks and intensities identified in the S2S data are publicly available for download under `ftp://s2sidx:s2sidx@acquisition.ecmwf.int/TCYC` and are based on the TC detection algorithm described in Vitart and Stockdale (2001). Accordingly, a candidate cyclone position is defined by the location of the closest local minimum in mean sea-level pressure surrounding a local maximum ($> 0.35\,\mathrm{s}^{-1}$) in 850-hPa relative vorticity. This position is considered to represent a TC, if local maxima in 200–500-hPa layer-averaged temperature and in 200–1000-hPa thickness occur within a radial distance of $2°$. From these maxima, temperature and thickness are further required to decline to all sides by at least $0.5\,\mathrm{K}$ and 50 m over $8°$, respectively. Vitart and Stockdale (2001) 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 ECMWF's 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 (see Section 4.1.2).

4.1.4 Tropical wave filtering

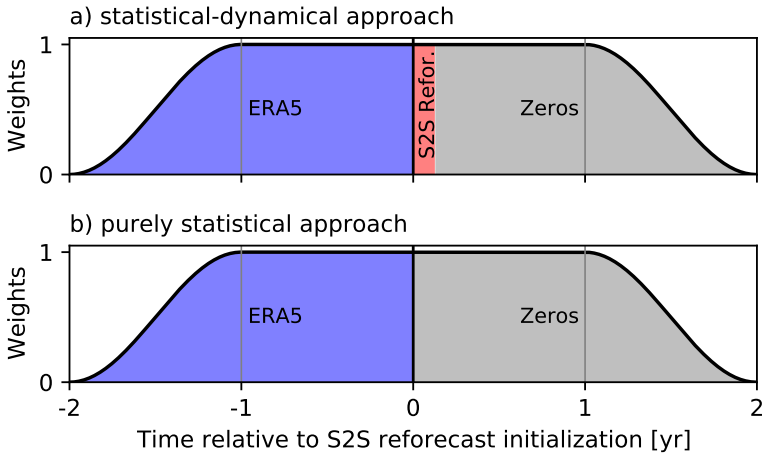

Figure 4.3: Illustration of how time series, filtered for tropical waves in real-time mode, are composed by different datasets, and how those are weighted over time in (a) the statistical-dynamical, and (b) the purely statistical approach, respectively. Reprinted from Maier-Gerber et al. (2021). © 2021, American Meteorological Society. Used with permission.

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. Wheeler and Kiladis (1999) suggest a filtering method, which applies a two-dimensional fast Fourier transform (FFT) along time and longitude. Comparing filtered observational data with theoretical solutions of the shallow water equations, Wheeler and Kiladis (1999) elaborate a set of filter windows in wavenumber–frequency domain, one window for each wave type. By setting to zero all signals outside of a certain window, and then passing it through an inverse FFT, the method yields the original variable filtered for a particular wave type. But since ringing artifacts typically occur at the edges of finite time series filtered in this way, Wheeler and Weickmann (2001) suggest to pad an array of zeros at the end of the time series to allow for real-time application of this method.

To create predictors representing tropical waves, the current study follows the concept of Janiga et al. (2018), who apply the wave filtering of Wheeler and Kiladis (1999) to a time series composed of reanalysis data and S2S reforecasts, with the zero-padding strategy of Wheeler and Weickmann (2001) used at the end. 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. 4.3a). 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. 4.3b). To more specifically evaluate subseasonal signals during filtering, the first four harmonics of the 1979–2018 annual cycle are calculated from

Table 4.1: Ranges of wave periods p and zonal wave numbers k used to filter for specific tropical wave modes. The 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	$2.5 : 10$	$-20 : 0$

ERA5 and subtracted from all non-zero portions of the composed time series. As illustrated in Figure 4.3, 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 each latitude and ensemble member separately, to filter in frequency–wavenumber domain for the five wave types listed in Table 4.1. 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, especially on subseasonal timescales. Although these windows were defined for solutions of the shallow water equations with the equator as the waveguide, the latitude-specific filtering is performed up to the midlatitudes for pragmatic reasons.

4.2 Isotonic distributional regression

Statistical post-processing is required to correct for systematic errors in TC predictions of the dynamical S2S model. Usually, this is done by fitting distributional regression models where the parameters of the predictive distribution are estimated from training data. A non-parametric alternative for statistical post-processing is based on IDR (Henzi et al., 2021). Based on the pool adjacent violators (PAV) algorithm, an approach to iteratively solve monotonic regression problems, IDR learns cumulative distribution functions (CDFs) conditioned on a set of covariances provided. During learning, the only constraint imposed is that the CDFs are isotonic (i.e., monotonically increasing) on the covariate space, which is realized by choosing and applying a partial order. Applying IDR requires the specification of a partial order on the covariate space. In this study, the S2S ensemble members are used as covariates which can be considered exchangeable. As discussed in Henzi et al. (2021), the increasing convex order is a suitable choice of partial order for

IDR in this setting for groups of exchangeable, real-valued covariates. A strength of the IDR method is evident from the fact that predicted distributions are already calibrated.

4.3 Validation strategy

A systematic comparison of different model approaches requires an appropriate strategy for validation. While forecasts from a climatological model can in principle be issued for any lead time outside the period used to generate the climatology, predictions of the target variables derived directly from the S2S-ECMWF model and the generation of S2S-based predictors for statistical-dynamical models depend on the twice-weekly produced reforecasts, thus posing the strongest constraint to a potential validation dataset in this study. Starting from each of the S2S reforecast initialization dates, for every model to be validated, forecasts are generated for the first five consecutive weeks, i.e. days 0–6, 7–13, 14–20, 21–27, 28–34. However, forecasts are only considered for validation if the middle of the respective forecast week falls into the North Atlantic hurricane season. This yields a total of 1040 validation instances (52 reforecasts per season $\times$ 20 seasons), for which S2S ECMWF reforecasts are available.

In contrast to the S2S-based model, the statistical models require a training dataset that is independent of the validation dataset. To fully exploit the relatively small number of S2S reforecasts for training and validation, 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 of each fold. The validation results are then averaged over all folds if not stated otherwise. Although forecasts are generated for every gridpoint and forecast week separately, the gridpoints within each of the two subregions are pooled to allow more solid conclusions to be drawn during model validation.

4.4 Forecast verification for tropical cyclone occurrence

Since all TC occurrence model output probabilities, predictions cannot be readily distinguished in right or wrong, as in the case of deterministic forecasts. Forecast verification therefore must assess the joint distributions of the binary target variable and the predicted probabilities. Focusing on different aspects of model validation, a variety of diagnostic concepts and tools have been developed in the literature to extract information from these high-dimensional joint distributions in a summarizing manner. To conduct a thorough validation of the TC occurrence models, we follow the usual strategy of applying several of these diagnostics (e.g., Gneiting and Vogel, 2019), and analyze a model's calibration, its potential and actual predictive skill, and the long-term expected costs for a user

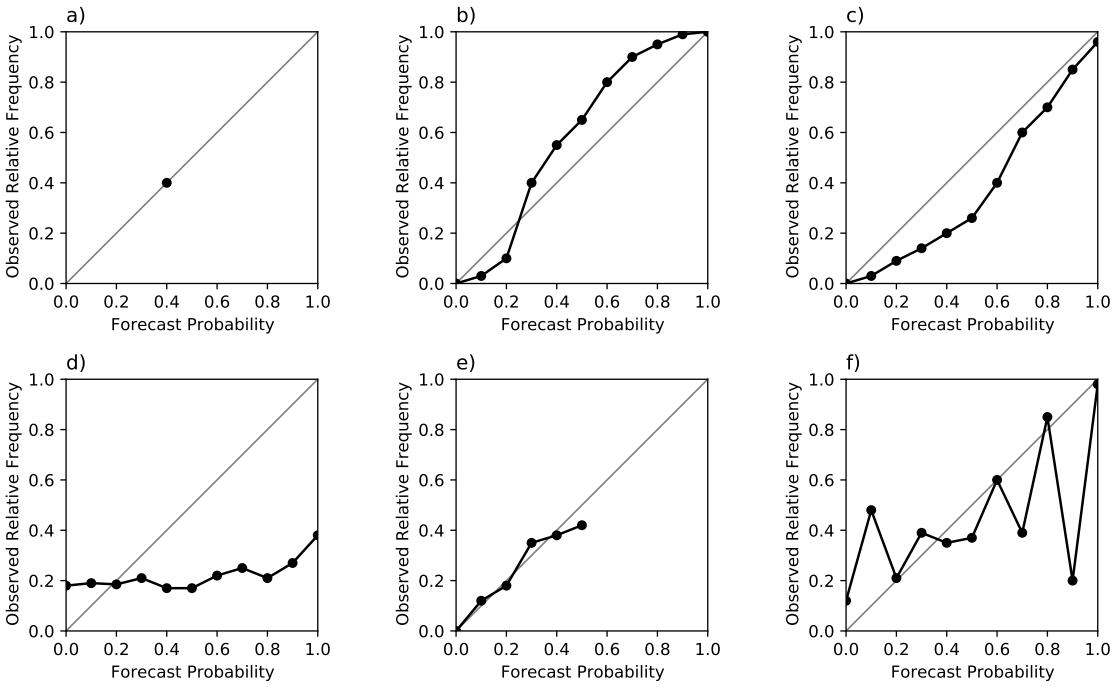

Figure 4.4: Reliability diagrams for different characteristic types of calibration curves. See text for a discussion.

when taking action based on a particular model. The individual tools and metrics used to address these aspects are introduced in the following.

4.4.1 Reliability diagram

A critical requirement for probabilistic forecasts is calibration. According to Gneiting et al. (2007), forecasts are calibrated if they are statistically consistent with their corresponding observations. We here follow the notion of probabilistic calibration (hereafter only referred to as 'calibration' for simplicity) introduced by Gneiting et al. (2007), and think of a set of N forecasts as being calibrated if the observed predictive distribution matches the forecast one for $lim_{N\to\infty}$. In case of probability forecasts for binary target variables, a common visual tool to assess calibration is the reliability diagram (Sanders, 1963; Wilks, 2011), which displays the joint distribution factorized into model reliability (calibration curve) and refinement (histogram). Calibration curves are obtained by conditioning the observed relative frequency on the forecast probability, and allow to identify unconditional and/or conditional biases. Miscalibration can thus be visually examined through deviations of the calibration curves from the diagonal. A model is said to be calibrated (or reliable) if the forecast probabilities match the observed relative frequencies. For example, if we issue 100 probability forecasts of 20%, the predicted event should occur 20 times.

Figure 4.4 illustrates characteristic calibration curves, some of which exhibit certain types of miscalibration. A model predicting a constant value yields a single point in the diagram, which is located on the diagonal in case of climatological forecasts (Fig. 4.4a). If the climatology allows for seasonal variations, the resultant calibration curve should follow the diagonal as climatological forecasts are expected to be well-calibrated, unless the climatology is subject to long-term trends. Unconditional biases are present if the calibration curve consistently deviates from the diagonal in one direction. A model has an underforecasting (overforecasting) bias if it generally predicts an event too seldom (often, Fig. 4.4c). In contrast, conditional biases are characterized by inconsistent deviations of the calibration curve from the diagonal. While an overconfident model tends to underforecast low probabilities and overforecast high probabilities (Fig. 4.4d), an underconfident model behaves the other way round (Fig. 4.4b). In case of rare events, forecast probabilities are generally rather low and the corresponding calibration curve may not span the full range of possible forecast probabilities (Fig. 4.4e). Another characteristic behaviour is an unstable calibration curve (Fig. 4.4f), which typically results from too few samples populating bins or a validation dataset being generally too small.

To reduce the dimensionality of the joint distribution, forecast-observation pairs are often grouped into and averaged within equally sized, uniformly distributed, or quantile-based bins, based on the forecast probability. However, the resulting calibration curve can be highly sensitive to the choice of the number of bins. Dimitriadis et al. (2021) solved this problem and proposed the so-called CORP approach, which has the advantage, among others, of providing optimally binned and readily reproducible diagrams. This optimal binning is achieved through the PAV algorithm. Here, we use the approach to avoid that a subjective choice for the number of bins may impact model validation.

4.4.2 Receiver operating characteristic curve

Receiver operating characteristic (ROC; Fawcett, 2006) curves are graphical tools for assessing the predictive ability of probabilistic forecasts in binary classification problems. Originally proposed in signal detection theory (Egan et al., 1961), ROC curves were first used for meteorological applications by Mason (1982), before they became increasingly popular from the late 1980s on (e.g., Swets, 1988). Using a threshold x, the probability forecast can be transformed into a hard classifier, by predicting an event if the probability forecast is $> x$ and no event if probability forecast is $<= x$. For any choice of threshold a two-by-two confusion matrix can be computed, defined in Table 4.2. From these elementary numbers, the hit rate (H) and false alarm rate (F) are calculated as follows:

$$H = \frac{a}{a+c} \quad \text{and} \quad F = \frac{b}{b+d}. \tag{4.2}$$

Table 4.2: Binary confusion matrix containing the relative frequencies a, b, c, and d of the four possible scenarios. Note that $a + b + c + d = 1$.

Actual class / Predicted class	Positive	Negative
Positive	True Positive a	False Positive b
Negative	False Negative c	True Negative d

The ROC curve plots H against F for all possible thresholds in the interval $[0, 1]$ where linear interpolation yields the final piecewise linear curve. A scalar measure that is widely used in conjunction with ROC curves is the area under the ROC curve (AUC), which expresses the model's predictive performance in terms of class separability. If a model is able to perfectly distinguish between classes for all probability thresholds, the ROC curve will connect the points (0,0), (0,1), and (1,1), and the AUC equals 1. In contrast, if a model outputs useless forecasts, the ROC curve follows the diagonal from (0,0) to (1,1), resulting in an AUC of 0.5. In this case, H and F are identical for all probability thresholds, meaning that the probability for true positive is the same as the probability for false positive. ROC curves running below the diagonal ($AUC < 0.5$) indicate that a model has skill but distinguishes classes in reverse manner, which, however, can potentially be exploited.

Properties of ROC curves, and hence of AUC, are that they are invariant under changes in proportions of the actual classes (Fawcett, 2006), and under strictly monotonic transformations applied to the forecast probabilities (Gneiting and Vogel, 2018). Given the definition of IDR (cf. Section 4.2), this second property implies that an IDR-based calibration does not alter the ROC curve of a model. Furthermore, ROC curves are insensitive to miscalibrated forecasts (Jolliffe and Stephenson, 2012; Wilks, 2011), and therefore should be used in concert with reliability diagrams for a comprehensive evaluation of model performance. This inability to reflect conditional and unconditional biases limits the use of AUC as it can merely assess a model's potential predictive skill. For final conclusions in terms of predictive skill, however, additional measures are needed.

4.4.3 Brier (skill) score and its decomposition

A performance measure that can assess calibration is the mean Brier score (BS; Brier, 1950). For a set of N forecast-observation pairs, it is defined as

$$BS = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{p}_i)^2, \tag{4.3}$$

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. Summing over quadratic forecast errors, the BS takes on values from 0 to 1, and is negatively oriented (i.e., lower is better). A perfect forecast thus yields a BS of 0, whereas a forecast predicting the opposite event results in a BS of 1. The BS is a strictly proper scoring rule, meaning that the expected score is uniquely minimized by the true underlying distribution of the observation (Gneiting and Raftery, 2007). This useful property guarantees that models with an honest assessment are rewarded, or to put differently, that miscalibrated models are penalized and thus cannot achieve the best score.

The reliability-refinement factorization of the joint probabilities, used to construct reliability diagrams (cf. Section 4.4.1), allows to decompose the BS into three components (Sanders, 1963; Murphy, 1973), assessing forecast reliability (REL), and resolution (RES), as well as the uncertainty of the target variable (UNC):

$$BS = \underbrace{\frac{1}{N}\sum_{k=1}^{K} N_k(\bar{\hat{p}}_k - \bar{y}_k)^2}_{\text{=REL}} - \underbrace{\frac{1}{N}\sum_{k=1}^{K} N_k(\bar{y}_k - \bar{y})^2}_{\text{=RES}} + \underbrace{\bar{y}(1+\bar{y})}_{\text{=UNC}}, \tag{4.4}$$

where K is the number of bins used to represent the range of forecast probabilities, and N_k is the number of forecasts populating the k-th bin. $\bar{\hat{p}}_k$ and $\bar{y}_k$ are the within-bin averaged forecast probabilities and observations, respectively, and $\bar{y} = \frac{1}{N}y_i$ is the observational sample climatology. Stephenson et al. (2008) note that, when different forecast probabilities are unified within bins, the decomposition in Eq. (4.4) requires two additional terms for within-bin variance and within-bin covariance, respectively (not shown). Because this frequently used decomposition strongly depends on the chosen binning strategy, and therefore cannot provide stable results, Dimitriadis et al. (2021) suggest a robust alternative. Based on the optimal binning generated by their CORP approach, the BS can be decomposed into terms for miscalibration (MCB), discrimination (DSC), and uncertainty (UNC):

$$BS = \underbrace{(BS - BS_{cal})}_{\text{=MCB}} - \underbrace{(BS_{clim} - BS_{cal})}_{\text{=DSC}} + \underbrace{BS_{clim}}_{\text{=UNC}}. \tag{4.5}$$

The two basic components of this decomposition are

$$BS_{cal} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widetilde{\hat{p}}_i)^2 \quad \text{and} \quad BS_{clim} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2, \tag{4.6}$$

which measure the BS for a (re)calibrated version of the forecasts $\widetilde{\hat{p}}_i$, and for a reference forecast predicting the observational sample climatology $\bar{y}$, respectively. Thus, the MCB term represents a measure for forecast miscalibration, whereas the DSC term assesses the ability of the (re)calibrated forecasts to discriminate between events better than a model forecasting the sample climatology. As opposed to MCB and DSC terms, UNC does not depend on forecast probabilities $\hat{p}_i$, but on y_i only, and is hence the same for all models and forecast weeks. With the BS being negatively oriented, and all three terms taking on values ≥ 0, according to Eq. (4.5), a good BS is achieved by a small contribution of the MCB term, and large contributions of the DSC term.

While the BS evaluates the performance of an individual model, the Brier skill score (BSS) provides information about skill gain relative to a reference model. This measure builds on the BS and puts into relation the score of the model to be validated (BS) with the score of a perfect model ($BS_{perf} = 0$), both as improvements from the score of a reference model BS_{ref}, in the following chosen as the mean seasonal climatology (MSC, see Section 6.1.1)

$$BSS = \frac{BS_{ref} - BS}{BS_{ref} - BS_{perf}} = \frac{BS_{MSC} - BS}{BS_{MSC} - 0} = 1 - \frac{BS}{BS_{MSC}}. \tag{4.7}$$

Using the MSC in this context has the advantage that the BS of the reference model is constant, and therefore allows a BSS-based comparison of models across lead times. Unlike the BS, the BSS is positively oriented, ranging from $-\infty$ to 1, but it is not strictly proper. A model to be validated has better, no, or less skill compared to the MSC model if the BSS becomes greater, equal, or less than zero, respectively.

4.4.4 Economic value

As informative as AUC and BS are for evaluating a model's predictive skill, they do not express how useful predictions are for a user in terms of actual costs. For this purpose, simple cost-loss decision models have been proposed in the meteorological literature (e.g., Ångström, 1922; Murphy, 1977), with the aim to assess economic value of forecast models for binary target variables. We here follow the approach of Richardson (2000, 2003), who define a value measure relative to a perfect forecast model. The basis for its derivation is laid by a decision model, which describes scenarios of costs C and losses L to a user, when acting based on a deterministic forecast. The scenarios result from

Table 4.3: Table of costs (C) and losses (L) associated with the four scenarios in Tab. 4.2, based on the cost-loss decision model described in the text.

Predicted class \ Actual class	Positive	Negative
Positive	C	C
Negative	L	0

the four possible combinations of predicted and actual classes (Tab. 4.3). In case of a positive prediction of the event, the user is thought to invest C to prevent from paying L, independent whether the event actually happens or not. In case of a negative prediction of the event, no action is taken by the user, and L will be lost if the event occurs, and 0 if not. The applicability of this decision model is subject to the assumption that C and L can be quantified, and that a user is interested in minimizing them in the long term, i.e. to consider average expenses.

For the construction of their relative measure, Richardson (2000, 2003) defines two reference forecasts. As the first reference, he considers expenses of a user, who has no short-term forecast at hand but acts on climatological information. The only possible reduction in climatological expenses results from the choice of either always investing C, or accepting to pay L in $\bar{y}$ of the cases:

$$E_{clim} = \min(C, \bar{y}L). \tag{4.8}$$

The second reference describes a user, who always makes the right decision, resulting in perfectly minimized expenses. This absolute limit of minimal expenses can be achieved by investing C only before the event actually occurs, i.e. in $\bar{y}$ of the cases:

$$E_{perf} = \bar{y}C. \tag{4.9}$$

The expenses for a deterministic forecast to be evaluated are obtained by multiplying the elements of the confusion matrix (Tab. 4.2) with the corresponding elements of the cost-loss table (Tab. 4.3), and adding them up:

$$E_{det} = (a + b)C + cL. \tag{4.10}$$

Combining these elementary scenarios of expenses, the economic value is defined as

$$V = \frac{E_{clim} - E_{det}}{E_{clim} - E_{perf}} = \frac{min(C/L, \bar{y}) - (a + b)C/L - c}{min(C/L, \bar{y}) - \bar{y}C/L}. \tag{4.11}$$

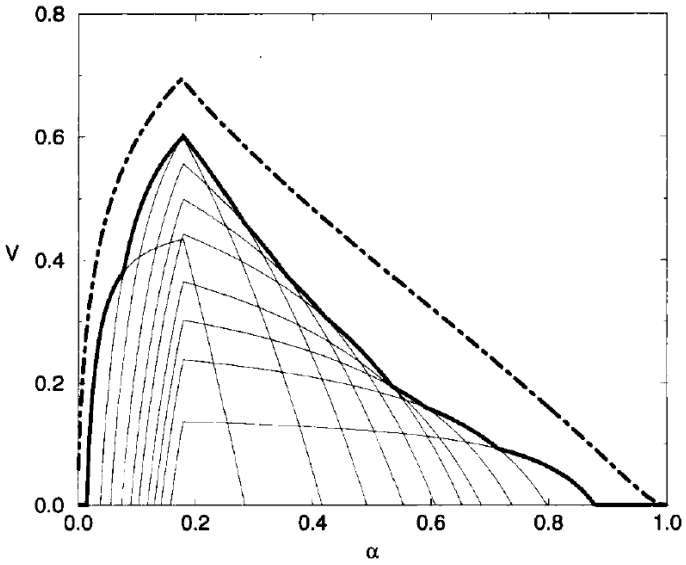

Figure 4.5: Economic value curves of an example ensemble prediction model for a set of probability thresholds (thin lines) and their envelope (thick solid line). Reprinted from Richardson (2000). © 2000 John Wiley & Sons, Inc. Used with permission.

Note that the right-hand side of the second equation is expanded by $1/L$, to formulate V as a function of the cost-loss ratio C/L. By subtracting E_{det} and E_{perf} from E_{clim}, respectively, the numerator and denominator become savings with respect to the climatological expenses. The economic value therefore expresses the fraction of the maximum achievable savings that can be obtained with a given deterministic model. V is positively oriented, i.e. higher is better, and usually plotted for values between 0 and 1, as this range indicates an economic benefit of the deterministic forecast information over the climatological information. Moreover, since for a user the investment of C is only profitable if $C/L < 1$, C/L is typically considered over the range from 0 to 1. For each model, its maximum value of $V_{max} = H - F$ is achieved for a cost-loss ratio of $C/L = \bar{y}$ (Richardson, 2003). Setting into ratio the economic values of any two models yields a ratio of their numerators and thus of the absolute savings that can be achieved with these models.

Since the TC occurrence models used in this thesis forecast probabilities, the elements of the confusion matrix, a, b, c and d, depend on the choice of a probability threshold p_{thres}. Once chosen, it turns a probabilistic forecast into a deterministic one. Because a, b and c are included in Eq. (4.11), the economic value curve also depends on p_{thres}, as illustrated in Fig. 4.5. Given a calibrated forecast model, however, Richardson (2000) states that its maximum economic value is reached when setting $p_{thres} = C/L$, i.e. using the user's cost-loss ratio. By applying this choice for a range of C/L-values, an envelope can be constructed, representing the overall maximum economic value achievable for the probabilistic forecast model being evaluated (Fig. 4.5).

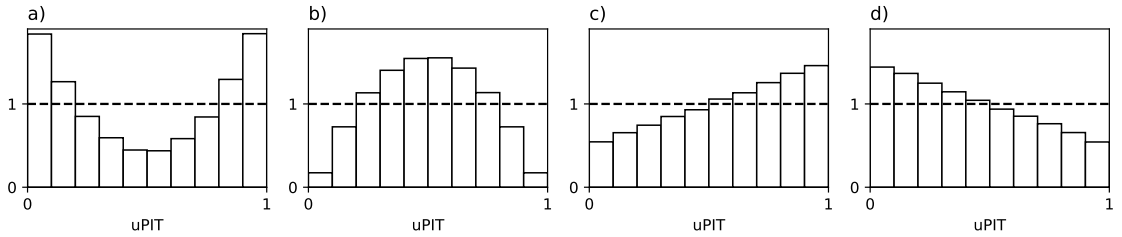

Figure 4.6: Illustration of typical (u)PIT histograms in case of an (a) overconfident, (b) underconfident, (c) underforecasting, and (d) overforecasting model, respectively. The dashed line highlights uniformity, which would characterize a calibrated model.

4.5 Forecast verification for accumulated cyclone energy

In contrast to the probability forecasts for TC occurrence, the models considered in the second part of this study output a full predictive distribution for the target variable ACE. Moreover, the target variable is no longer binary but real-valued. This leads to a much more complex situation in terms of forecast verification, and therefore fewer concepts and tools for model validation are available in the literature. Nevertheless, and in analogy to the tools used for validating the TC occurrence models, there are tools to examine and assess a model's calibration, as well as its potential and actual predictive skill. The individual tools and metrics used to address these aspects are introduced in the following subsections.

4.5.1 Unified probability integral transform histograms

While predicting full distributions allows for an assessment of forecast uncertainty in the first place, it raises the question of how statistical consistency with observations (i.e., calibration) should be conceived when verifying distributions. A useful tool that allows to visually analyze model calibration builds on the probability integral transform (PIT) theorem, which states that if the CDF of any random variable is in turn considered as another random variable, the resulting distribution is standard uniform. Let Y denote the observational continuous random variable, and $\hat{P}$ is the forecast CDF. Then the set of N forecasts is calibrated if the corresponding PIT values $\hat{P}(Y)$ are standard uniformly distributed. Because the probability of observing no ACE is typically greater than zero, $\hat{P}$ shows a discontinuity at ACE $= 0$ and exhibits further discontinuities for ACE > 0 when the CDF is expressed nonparametrically, as is the case with the IDR method, for example. Therefore, the PIT value at each discontinuity is $\hat{P}(Y-)+V(\hat{P}(Y)-\hat{P}(Y-))$, using a standard uniform distribution V to randomly draw a value between the left-hand and right-hand limits.

When a set of PIT values is plotted in a histogram, usually referred to as PIT histogram (e.g, Gneiting et al., 2005), any type of deviation from the expected uniformity (i.e., from a flat histogram) indicates a certain type of miscalibration. An U-shaped PIT histogram reveals an overconfident forecast distribution, i.e. the observations materialize too often beyond the predicted extremes (Fig. 4.6a). In contrast, an inverse U-shape hints at an underconfident forecast distribution, where the observed values are usually less extreme than expressed by the predicted uncertainty (Fig. 4.6b). In addition to the identification of confidence biases resulting from misrepresented forecast uncertainty, a PIT histogram also conveys information about unconditional biases. If a PIT histogram is skewed to the left, the median is greater than the mean, characterizing an underforecasting bias (Fig. 4.6c). Consequently, the opposite case is referred to as an overforecasting bias (Fig. 4.6d).

In case the forecast distribution is composed of a finite number of forecast values, e.g. from ensemble members or past observations, rank histograms (Anderson, 1996; Talagrand et al., 1997; Hamill and Colucci, 1997) are the analogon for such an analysis. Given a set of m forecast values f_{ij} for the i-th forecast instance, $j = 1, \ldots, m$, the verification rank r_i indicates the rank of the observed value y_i with respect to the sorted set of f_{ij}, taking on values between 1 and $m + 1$. In case the observation is tied with some forecast values, the rank is determined as a random draw between the smallest and largest possible ranks.

Since rank histograms can be interpreted in the same way as PIT histograms (Hamill, 2001), Vogel et al. (2018) suggest to unify both types, allowing for a comparison between continuous and discrete forecast distributions. To achieve this, unified PIT (uPIT) values are chosen to be equivalent to conventional PIT values in case of continuous distributions, while the verification ranks are transformed into uPIT values in the discrete case by $(r_i - 1+U)/(m+1)$, where U is a standard uniform distribution. In this thesis, uPIT histograms are analyzed to assess calibration of the ACE models, due to the fact that both continuous and discrete predictive distributions are compared. As in Vogel et al. (2018), a fixed number of 20 equally sized bins is used to display uPIT histograms in a consistent manner.

4.5.2 Universal ROC curve and coefficient of predictive ability

A considerable limitation of the concept of ROC curves is that it is only applicable to binary target variables (cf. Section 4.4.2), such as TC occurrence, and thus cannot be readily used for ACE as the target variable. For real-valued point forecasts, however, Gneiting and Walz (2019) propose universal ROC curves (UROC) as a generalization to real-valued target variables. Given a validation dataset of N forecast-observation pairs, the distinct real-valued realizations of the target variable are referred to as classes of

which there is a finite number $m \leqslant N$. By successively thresholding the target variable at each class value, $m-1$ binary classification problems are obtained, on which a classical ROC curve can be computed. A UROC curve combines the $m-1$ corresponding individual ROC curves in form of a weighted average

$$UROC = \sum_{c=1}^{m-1} w_c ROC_c, \qquad c = 1, \ldots, m-1, \tag{4.12}$$

where ROC_c is the ROC curve, and w_c the corresponding weight of class c, respectively. The class-specific weights are defined as follows, based on the numbers of instances n_i and n_j populating the consecutive classes i and j, respectively.

$$w_c = \left(\sum_{i=1}^{c} n_i \sum_{i=c+1}^{m} n_i \right) \Bigg/ \left(\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (j-i) n_i n_j \right), \tag{4.13}$$

The major advantage of UROC curves is that they represent an evaluation of the entire set of individual ROC curves as a whole, and therefore no choice of any threshold with respect to the target variable is required. To provide a measure for potential predictive skill, Gneiting and Walz (2019) also derive a real-valued equivalent to the binary-case AUC, which is referred to as coefficient of predictive ability (CPA), given by

$$CPA = \sum_{c=1}^{m-1} w_c AUC_c. \tag{4.14}$$

Like UROC curves, CPA is designed in form of a weighted average over the individual class-specific AUC values, and is equivalent to the area under the UROC curve. For this reason, its interpretation in terms of predictive power is analogous to the one of AUC for the binary case (cf. Section 4.4.2). Because ROC curves are invariant under strictly monotonic transformations, so is the UROC curve and associated CPA, which further implies that an IDR-based statistical post-processing will not affect UROC analysis.

4.5.3 Continuous ranked probability (skill) score

Since the forecast model output for ACE takes the form of a predictive cumulative distribution function, a score is needed that considers the following two aspects, stated in Gneiting et al. (2007), when assessing predictive performance. Firstly, the score should inform about how well a model is calibrated. And secondly, the score should measure the sharpness of the predictive distribution. To achieve this, the mean BS, used to assess calibration in case of two-category variables, can be extended to the contiguous multi-category case, resulting in the so-called mean ranked probability score (RPS; Epstein,

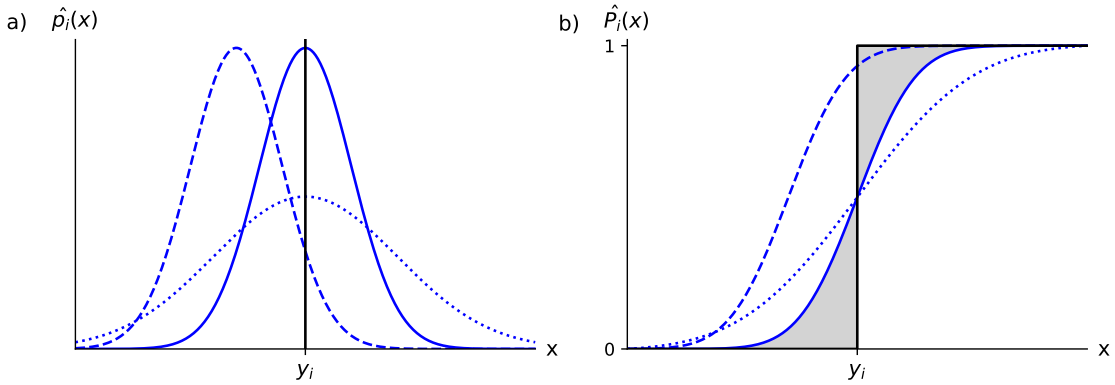

Figure 4.7: Examples of (a) probability density functions and observation, and (b) corresponding CDFs. The gray shaded area highlights the deviation of one of the predicted CDFs (solid blue) from the observational CDF (black), which is measured by the CRPS.

1969; Murphy, 1971).

$$RPS = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{M}\left(Y_{ij} - \hat{P}_{ij}\right)^2\right), \tag{4.15}$$

where N is the number of forecast-observation pairs, M is the number of categories, and Y_{ij} and $\hat{P}_{ij}$ contain the observational and forecast cumulative probabilities of the i-th instance, respectively. Because ACE is a continuous target variable, the multi-category case needs to be further extended to the continuous case to be applicable in the context of this thesis. The required infinitesimal extension turns the formulation of the mean RPS into the mean continuous ranked probability score (CRPS; Matheson and Winkler, 1976; Hersbach, 2000).

$$CRPS = \frac{1}{N}\sum_{i=1}^{N}\left(\int_{-\infty}^{\infty}\left(\hat{P}_i(x) - \mathbb{1}\{x \geq y_i\}\right)^2 dx\right), \tag{4.16}$$

where $\hat{P}_i$ is the predicted cumulative distribution function of the i-th instance, and y_i the corresponding observation.

A visual depiction of what Eq. (4.16) measures is provided by Figure 4.7. For a given forecast probability distribution (Fig. 4.7a), the CRPS integrates over the area between the corresponding cumulative distribution function and the step function located at the observation (Fig. 4.7b). Therefore, a forecast distribution should resemble the step function as close as possible, i.e. the probabilities should be concentrated (sharpness aspect) around the observed value (calibration aspect). From this it can be deduced that the CRPS is negatively oriented with values ranging from 0 to ∞. The CRPS has the unit of the underlying random variable, which in the case of ACE is $10^4 kt^2$, and is sensitive to

distance. Like the BS, the CRPS is a strictly proper scoring rule (Matheson and Winkler, 1976), i.e. in expectation the lowest score is achieved by predicting the true underlying distribution (Gneiting and Raftery, 2007). In case of a deterministic forecast, the CRPS reduces to the mean absolute error, which allows to directly compare with probabilistic forecasts.

As for the BS, a skill score can also be calculated for the CRPS to compare predictive performance between models. Following the usual definition of a skill score, i.e. setting into relation the score of the model to be validated ($CRPS$) with the score of a perfect model ($CRPS_{perf} = 0$), both as deviations from the score of a reference model ($CRPS_{ref}$), the continuous ranked probability skill score (CRPSS) is defined as

$$CRPSS = \frac{CRPS_{ref} - CRPS}{CRPS_{ref} - CRPS_{perf}} = 1 - \frac{CRPS}{CRPS_{FSPC}}. \tag{4.17}$$

Note that the right-hand side of the second equation results from using the full-season probabilistic climatology (FSPC, see Section 7.1.2) as the reference model. The FSPC is chosen here to be consistent with the MSC model used for the BS, as they are united in their approach of considering past observations from the entire hurricane season to compose the underlying climatology. Since forecasts of the FSPC model are independent of the forecast week, so are its corresponding CRPS values. Hence, CRPSS values for a model to be validated can be readily compared across lead times. As opposed to the CRPS, the CRPSS is positively oriented with values ranging from $-\infty$ to 1, but it is not strictly proper. A model to be validated has better, no, or less skill compared to the FSPC model if the CRPSS becomes greater, equal, or less than zero, respectively.

5. 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, 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.

5.1 Statistical-dynamical predictors

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. When these are provided as separate predictors, the statistical models should learn primarily from the predictive signals associated with the ensemble means, as long as 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 4.1.1. 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 the ensemble standard deviation.

5.1.1 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 5.1a 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. 5.1b). The area with the highest correlations of greater 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 North Atlantic TC activity (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. 5.1c) and mean MDR SST (Fig. 5.1b) underpins the ENSO teleconnection effect on MDR environmental conditions, identified by Gray (1984), and thus the potential value for including this remote predictor type.

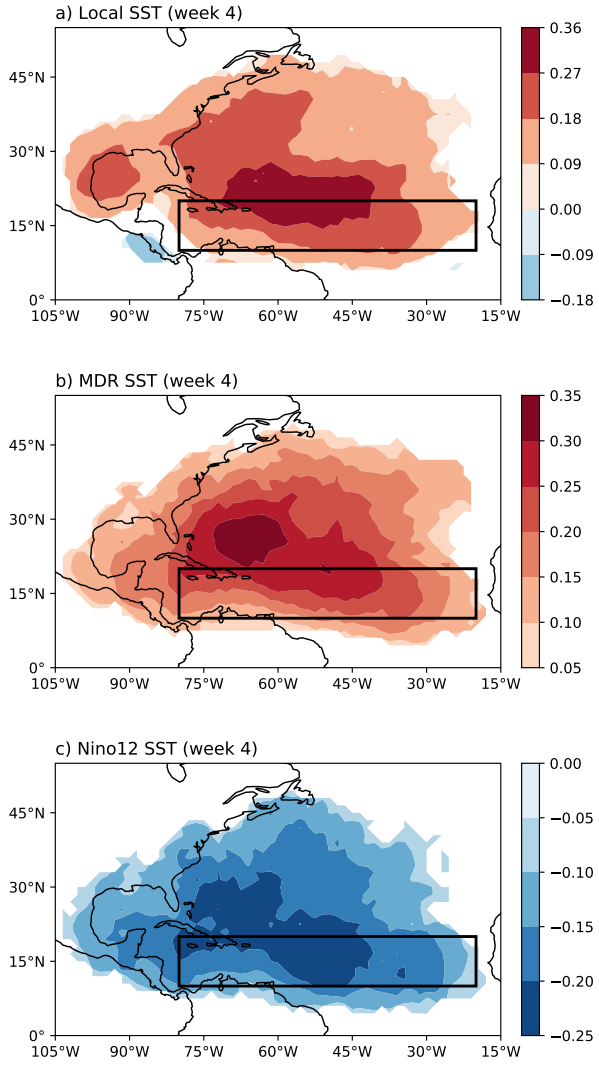

Figure 5.1: Predictor–target Pearson correlation coefficient for ensemble mean (a) local SST, (b) MDR SST, and (c) Niño1+2 SST at week four. Values are only displayed where correlations are statistically significant at a significance level of 5%. The thick black line highlights the MDR. Note the varying ranges of different color bars. Reprinted from Maier-Gerber et al. (2021). © 2021, American Meteorological Society. Used with permission.

5.1.2 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 (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-hPa to 850-hPa vertical shear, and potential intensity. The latter, however, could not be calculated from the S2S database.

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. 5.2a). 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 local maximum in the western Gulf of Mexico and a pronounced maximum west of the West African coast (Fig. 5.2b). 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. 5.2c).

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 filter for tropical wave modes as described in Section 4.1.4. Although tropical waves are tied to the equator, their influence can be attributed to TC formation at latitudes beyond $30°\text{N}$ (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 $850\,\text{hPa}$ due to higher correlation coefficients. The resulting predictor–target correlation coefficients are predominantly negative for most wave types (Fig. 5.3a-d), meaning that a reduced upper-level wave activity facilitates TC occurrence. The positive

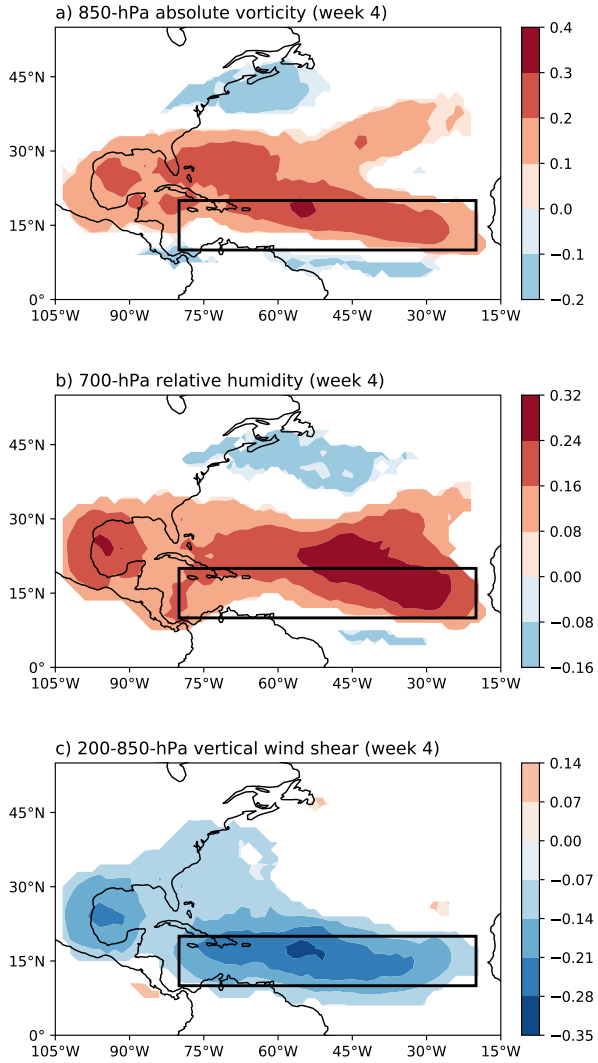

Figure 5.2: As Fig. 5.1, but for ensemble mean (a) 850-hPa absolute vorticity, (b) 700-hPa relative humidity, and (c) 200-hPa to 850-hPa wind shear at week four. Note the varying ranges of different color bars. Reprinted from Maier-Gerber et al. (2021). © 2021, American Meteorological Society. Used with permission.

correlations associated with Mixed Rossby-gravity/tropical depression (MRG/TD) waves (Fig. 5.3e) 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 subseasonal TC occurrence before (e.g., Leroy and Wheeler, 2008; Henderson 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[1] to define MJO remote predictors. RMM indices are often used to distinguish between eight circumglobal phases of

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

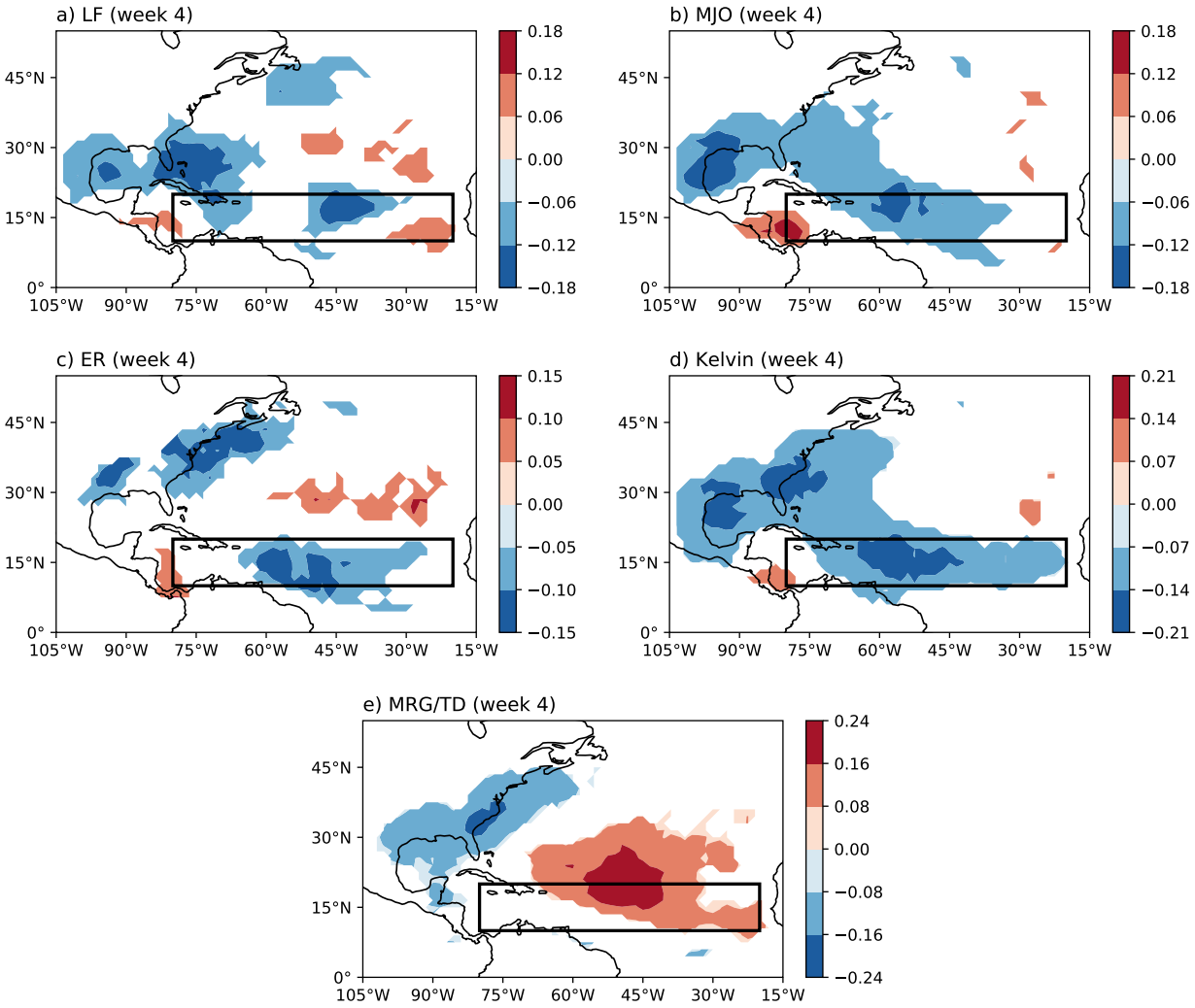

Figure 5.3: As Fig. 5.1, but for ensemble mean 200-hPa divergence squared at week four filtered for the wave types listed in Table 4.1. Note the varying ranges of different color bars. Reprinted from Maier-Gerber et al. (2021). © 2021, American Meteorological Society. Used with permission.

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.

5.1.3 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, especially in the MDR (Zhang et al., 2016, 2017a; 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, integrating over frequency and area of the identified PV streamers, correlates better with 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-hPa to 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. 5.4, 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. (2017a).

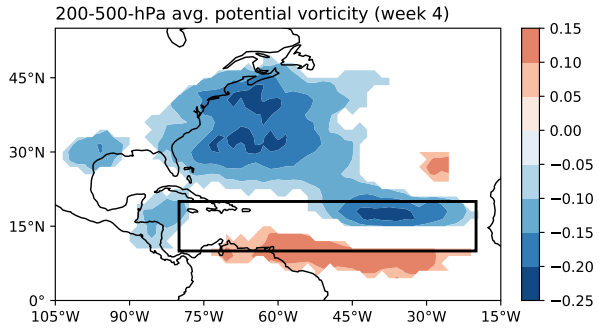

Figure 5.4: As Fig. 5.1, but for ensemble mean 200-hPa to 500-hPa layer-averaged PV at week four. Reprinted from Maier-Gerber et al. (2021). © 2021, American Meteorological Society. Used with permission.

5.2 Statistical predictors

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 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 on which the S2S reforecast was initialized.

6. Subseasonal forecasting of tropical cyclone occurrence

6.1 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.

6.1.1 Climatological forecasts

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. 4.1a, orange 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). Because the climatological models 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.

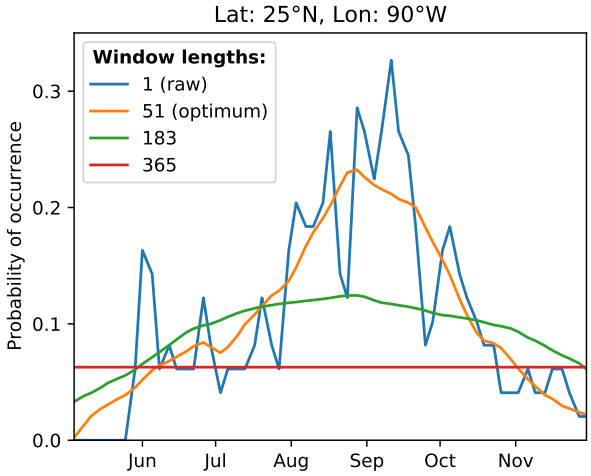

Figure 6.1: Climatological seasonal cycle (CSC) at $25°$N and $90°$W exemplarily smoothed with uniform kernels of window length 1 (blue), 51 (orange), 183 (green), and 365 (red) days, respectively. At this grid point, the CSC smoothed with a window length of 51 days was identified to best correlate with the target variable. Reprinted from Maier-Gerber et al. (2021). © 2021, American Meteorological Society. Used with permission.

Seasonal fluctuations evident from the CSC example in Fig. 6.1 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. When averaged over the gridpoints within the respective validation subregion, the optimal window length in the Gulf of Mexico turns out to be 48 days, whereas it is 24 days in the central MDR, thus differing by a factor of two. This indicates that more climatological data is needed in the Gulf of Mexico to build a robust climatological model that best represents the seasonal variations.

Given the chaotic nature of the atmosphere, the skill of any model should converge to the skill of the best performing climatological model for long enough lead times. For this reason, the CSCopt model 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-safe”, 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 subseasonal lead times (Scheuerer

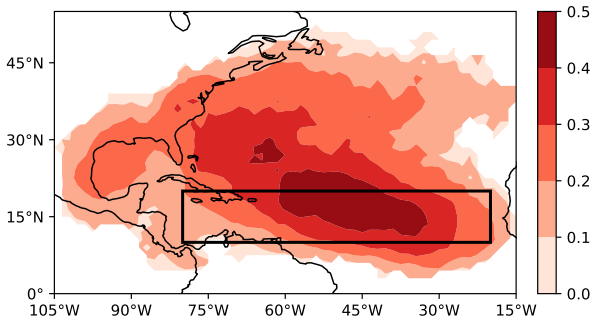

Figure 6.2: Predictor–target Pearson correlation coefficient for the locally optimized climatology. Values are only displayed where correlations are statistically significant at a significance level of 5%. The thick black line highlights the MDR. Reprinted from Maier-Gerber et al. (2021). © 2021, American Meteorological Society. Used with permission.

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 6.2 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. 4.1, red contour). Peak correlation values of up to 0.5 are located slightly north of the center of the MDR, 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 pool.

6.1.2 Dynamical forecasts

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 0000 UTC instances of the TC tracks identified in the S2S ECMWF 1998-2017 ensemble reforecasts (hereafter referred to as *S2STC*). However, Lee et al. (2018) point out that the combination of the inability of all S2S models to sufficiently resolve the TC inner-core thermodynamics, and the coarse resolution used for archiving the model output results in a systematic underestimation of the peak TC intensities. To take account of this intensity bias, they make a quantile-based comparison of the maximum intensities reached during the TCs' lifetimes between observed and forecasted storms. For the S2S ECMWF model,

they propose a lowered threshold of $24\,\mathrm{kn}$ instead of $34\,\mathrm{kn}$ to define the lower limit for tropical storm strength. Although earlier model version dates are used in Lee et al. (2018) than in the present study, a similar analysis for the reforecasts used here confirms this alternative threshold, and it is thus used to define TC occurrence for all S2S benchmark models.

Since S2STC forecasts are frequently not calibrated, we have tested different techniques to correct for potential miscalibration. The IDR method introduced in Section 4.2 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.[1] 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 6.3.1.

6.2 Statistical model development

6.2.1 Logistic regression

If the target variable is binary, 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

$$\hat{p_i}\left(\beta_0, \boldsymbol{\beta}\right) = logit^{-1}\left(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}\right) = \frac{1}{1 + \exp\left(-\beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)}, \tag{6.1}$$

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:

$$\min_{\beta_0, \boldsymbol{\beta}} \quad \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} - \sum_{i=1}^{N}\left(y_i \log\left(\hat{p_i}\right) + \left(1 - y_i\right)\log\left(1 - \hat{p_i}\right)\right). \tag{6.2}$$

[1] 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).

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		●	●	●

Figure 6.3: 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 provided to the statistical-dynamical and the purely statistical model, respectively. While dots denote predictors that can be chosen by the sequential predictor selection, ticks signify fixed predictors. Reprinted from Maier-Gerber et al. (2021). © 2021, American Meteorological Society. Used with permission.

The second term corresponds to maximum likelihood estimation, the first to an 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.

6.2.2 Sequential predictor selection

Training a logistic regression model on the full variety of predictors developed and motivated in Chapter 5 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 within each subregion to make predictor selection and model training more robust. An overview of the potential predictor pools, from which the two approaches can choose, is presented in Fig. 6.3. 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 fivefold CV on the training period. For a logistic regression model with P predictors, the AIC is defined as

$$AIC = -2\frac{LL}{N} + 2\frac{P}{N}, \tag{6.3}$$

where LL is the binomial log-likelihood based on N forecasts and 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 Henderson and Maloney (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.

Figure 6.4 shows the results of the sequential predictor selection applied to the statistical-dynamical approach. From the pool of 25 predictors provided to the statistical-dynamical approach, 15, 11, 10, 6, and 6 predictors were selected in the Gulf of Mexico for forecast week 1, 2, 3, 4, and 5, respectively (Fig. 6.4a). In analogy, the optimal subsets identified for the central MDR are comprised by 12, 6, 6, 7, and 8 predictors (Fig. 6.4b), thus using less predictors for the first three forecast weeks. In addition to the climatological seasonal cycle, predictors selected at all forecast weeks are mean 850-hPa absolute vorticity and mean 200-850-hPa vertical wind shear in the central MDR, and standard deviation of 850-hPa absolute vorticity in the Gulf of Mexico. In contrast, predictors that were not picked at all are standard deviation of low-frequency modes, mean MJO modes, standard deviation of MRG/TD waves, and mean Nino12 SST for both subregions, as well as mean low-frequency modes, standard deviations of ER waves, 200-500-hPa PV, and MDR SST, and mean Nino12 SST in the central MDR.

A comparison of the subsets over lead time reveals for the Gulf of Mexico that the majority of tropical wave, GPI, and extratropical predictors is primarily selected up to forecast week three, but no longer beyond (Fig. 6.4a), except for standard deviations of Kelvin waves and 850-hPa absolute vorticity, and mean and standard deviation of

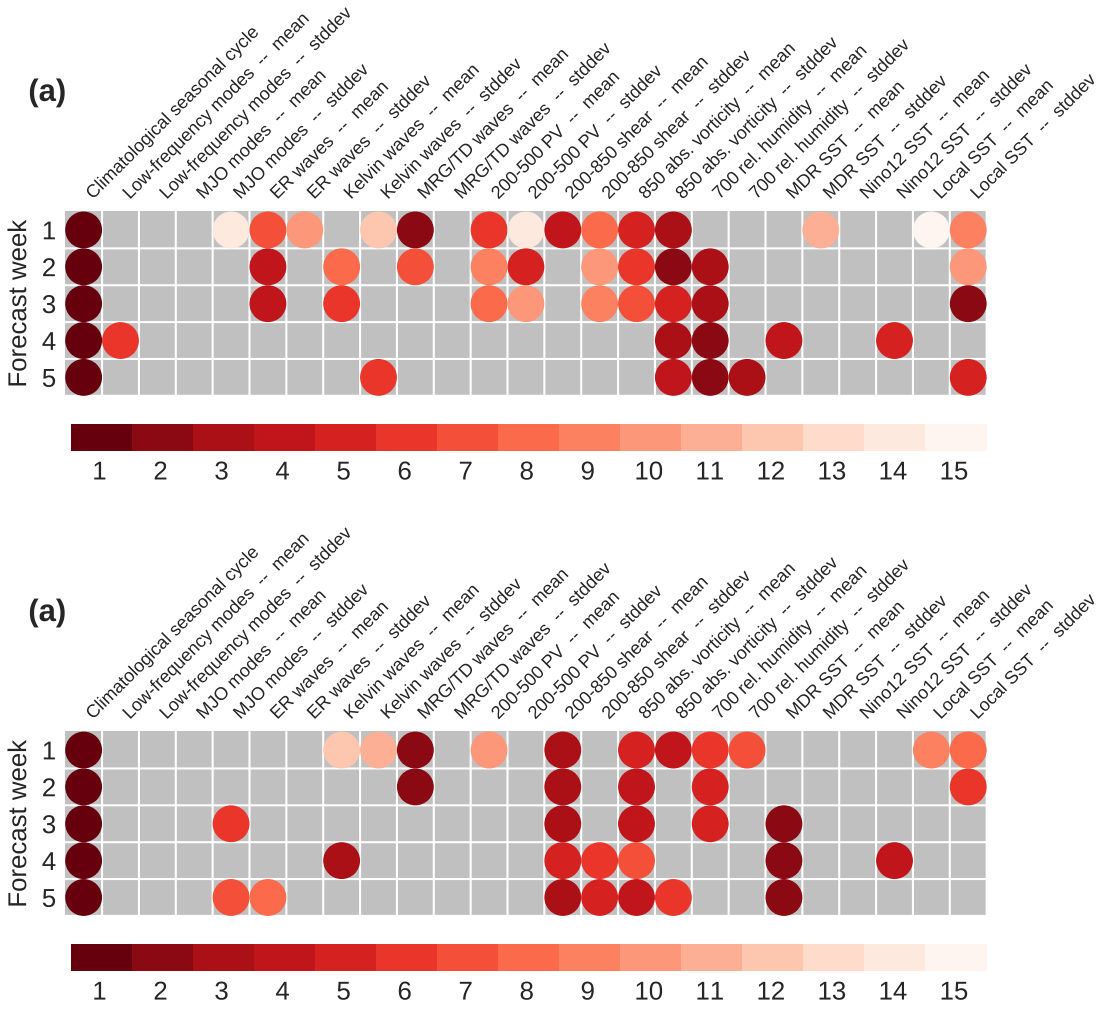

Figure 6.4: Subsets selected by the sequential predictor selection scheme for the statistical-dynamical approach for week one to five in (a) the Gulf of Mexico, and (b) the central MDR, respectively. Colors indicate the rank of the selected predictor averaged over the 20 folds of the cross-validation. If no circle is plotted, the predictor was not chosen. Reprinted from Maier-Gerber et al. (2021), supplementary material. © 2021, American Meteorological Society. Used with permission.

700-hPa relative humidity. In the central MDR, a striking transition is found within the oceanic predictor group between forecast week two and three (Fig. 6.4b). Mean MDR SST is chosen as the second predictor from week three on, whereas the mean and standard deviation of local SST were used on the medium range. Furthermore, a shift from higher-frequency (Kelvin and MRG/TD) to lower-frequency tropical wave modes (MJO and ER) occurs with increasing leadtime for the selected subsets in the central MDR.

Selection results for the purely statistical approach presented in Fig. 6.5 show that the optimal subsets in the central MDR are comprised by an equal or higher number of predictors compared to the Gulf of Mexico. From the pool of 13 predictors provided, 6, 4, 2, 3, and 3 (8, 4, 4, 4, and 3) predictors were selected in the Gulf of Mexico (central

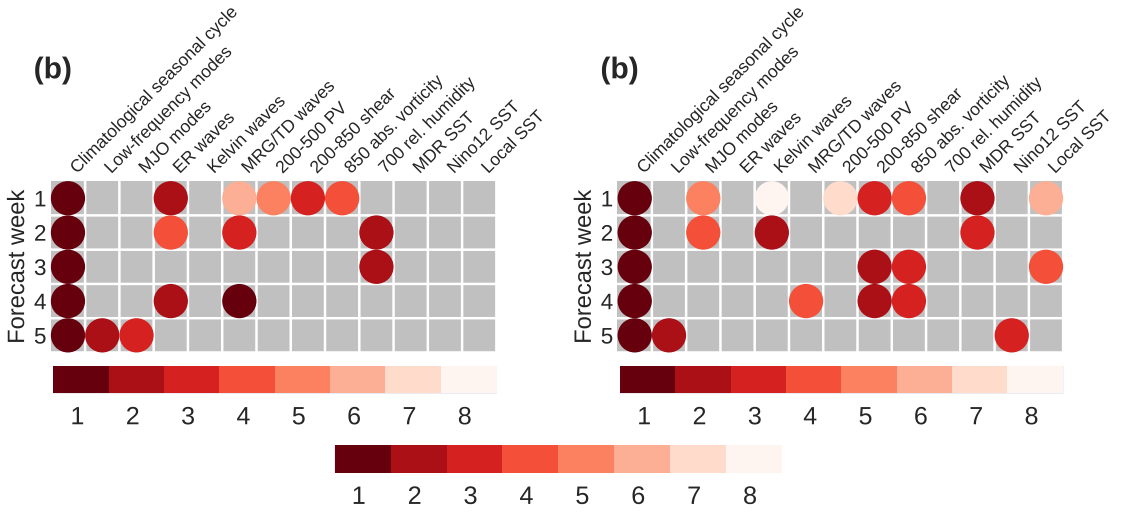

Figure 6.5: As in Fig. 6.4, but for the purely statistical approach. Reprinted from Maier-Gerber et al. (2021), supplementary material. © 2021, American Meteorological Society. Used with permission.

MDR) for forecast week 1, 2, 3, 4, and 5, respectively (Fig. 6.5a,b). Except for the climatological seasonal cycle, there was no predictor selected at all five forecast weeks in either subregion. However, consistency over leadtime appears by the non-consideration of all oceanic predictors and the Kelvin waves in the Gulf of Mexico, and of the ER waves and 700-hPa relative humidity predictors in the central MDR. When comparing results over the forecast week, only tropical wave predictors were chosen in the Gulf of Mexico beyond week three, whereas GPI and oceanic predictors are still included in the central MDR at week four and five, respectively.

6.3 Model comparison

6.3.1 Calibration

Figure 6.6 shows CORP reliability diagrams for the Gulf of Mexico and central MDR week-four forecasts to represent the subseasonal 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. 4.1b). 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

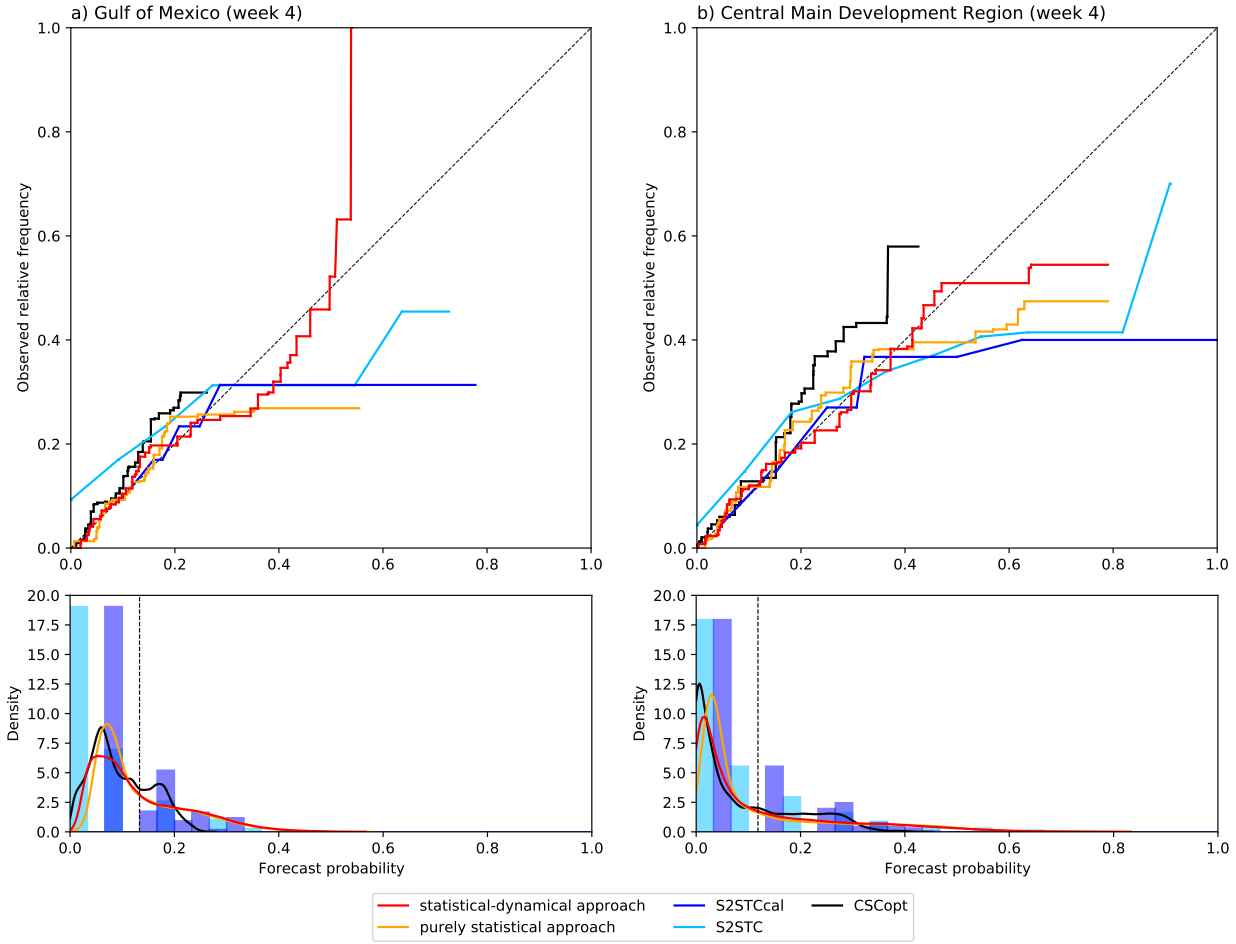

Figure 6.6: CORP reliability diagram for (a) Gulf of Mexico and (b) central MDR week-four forecasts, respectively. While forecast probability distributions are visualized by means of histograms for the S2STC and S2STCcal models, a kernel density estimation is applied to generate continuous curves for the other models. The dashed vertical line indicates the mean relative frequency of the target variable. Reprinted from Maier-Gerber et al. (2021). © 2021, American Meteorological Society. Used with permission.

the logistic regression models can predict with slightly higher confidence compared to the benchmark models.

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 4.2. 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, subseasonal 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.

6.3.2 Potential predictive skill

Potential skill of the TC occurrence models is assessed by means of ROC curves, which are displayed for forecast week one and four in Fig. 6.7. ROC curves for the remaining forecast weeks can be found in the appendix (Fig. A.1.5). Note that because of the invariance of ROC analysis under strictly monotonic transformation, ROC curves and AUC values for the IDR-calibrated S2STCcal model are identical to those for the S2STC model and are therefore not discussed separately. For both weeks and subregions presented, it is apparent that all models perform better than a random classifier model, as all AUC values are well above 0.5.

At week one, the statistical-dynamical model outperforms all other models, reaching an AUC of 0.83 in the Gulf of Mexico and 0.87 in the central MDR (Fig. 6.7a+b). Although the AUC value of the S2STC model is only 0.05 worse in both subregions, its potential predictive skill is comparable to the one of the CSCopt and purely statistical models in the central MDR, while it still exceeds them by about $+0.1$ in the Gulf of Mexico. This highlights that TC occurrence prediction for week one in subregions with a less pronounced seasonal cycle (here represented by the Gulf of Mexico) should either use a dynamical or statistical-dynamical modelling approach, with a slight preference for the latter. The purely statistical model virtually behaves like the CSCopt model for both subregions and forecast weeks. A comparison of AUC values for the CSCopt model between the two subregions, which is independent of the forecast week, reveals that climatological information has a much greater potential for predicting TC occurrence in the central MDR than in the Gulf of Mexico (0.83 vs. 0.68).

Using this model as reference, a decrease in skill is found for the S2STC and the statistical-dynamical models from week one to four. While the S2STC ROC curve clearly drops below the ones of the other models, by ≥ 0.7 in the Gulf of Mexico and ≥ 0.9 in the central MDR, respectively, the statistical-dynamical ROC curve merely falls back to CSCopt model (Fig. 6.7c+d). From visual congruence of ROC curves, but also from

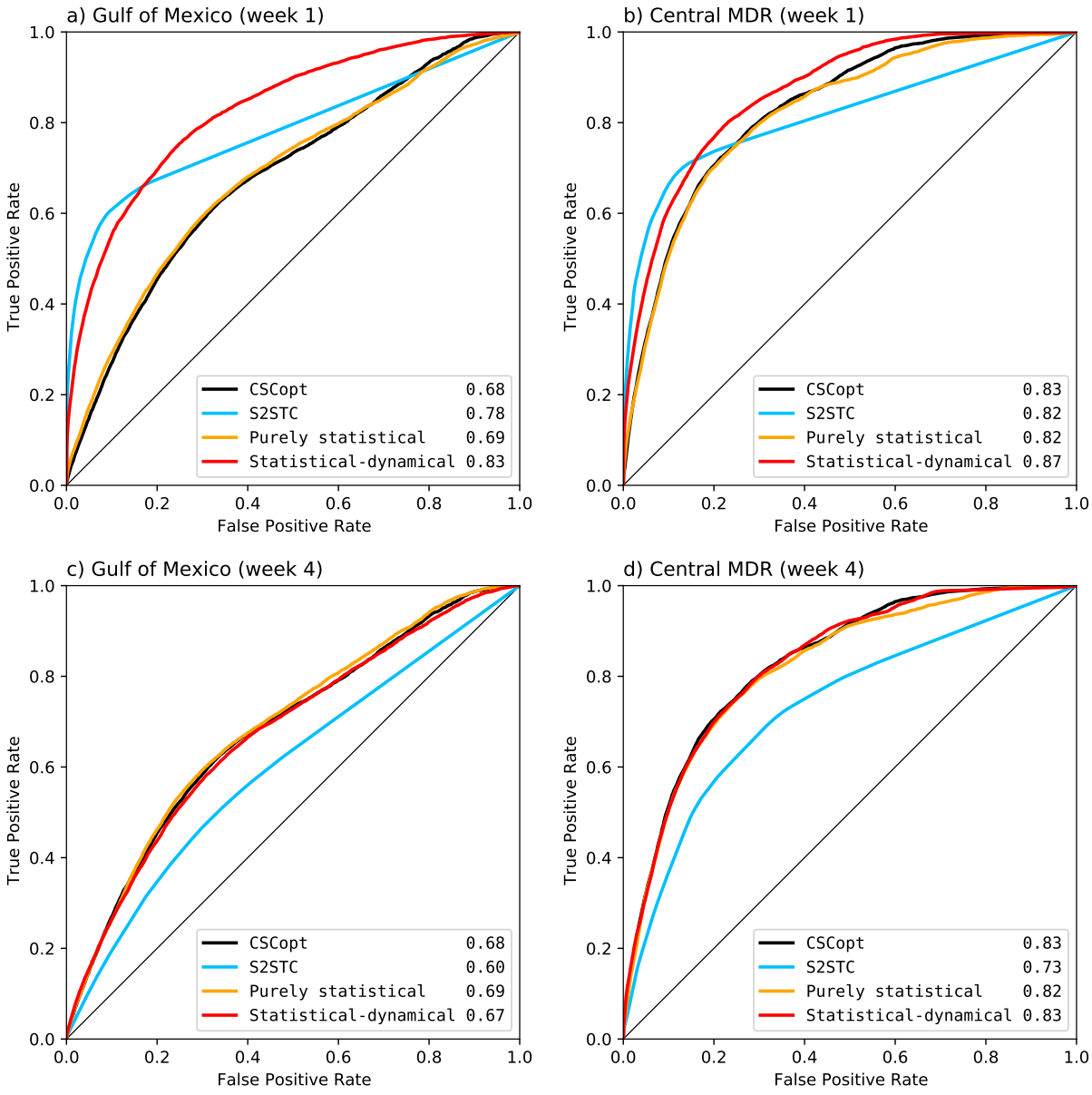

Figure 6.7: ROC curves for week-one (a+b) and week-four (c+d) forecasts of the CSCopt (black), the S2STC (lightblue), the purely statistical (orange), and the statistical–dynamical (red) model, respectively, in the Gulf of Mexico (a+c) and central MDR (b+d) subregions. The numbers in the legend of each panel show the model-specific area under the curve (AUC) averaged over the 20 folds.

corresponding AUC values, it appears that the CSCopt, the purely statistical, and the statistical-dynamical models perform comparably on subseasonal time scale. Therefore, from a practical point of view, one could argue that the CSCopt model is the preferred model type for this purpose, since it is the least complex model among these three and predictions can be reused without any effort after its one-time generation.

6.3.3 Actual predictive skill

Taking into account the aspect of calibration when validating models, Figure 6.8 shows the BSS as a function of forecast week validated in the Gulf of Mexico and central MDR subregions, allowing for comparison of actual predictive skill between models. Since climatological forecasts are independent of the forecast week, their 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 the CSC BSS being three times higher (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). CSC skill is further enhanced when correcting for the undersampling problem of the CSC model through a locally optimized smoothing (see Section 6.1.1 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 seasonal cycle. An optimal window length that is 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 model, and hence a greater potential for 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 BSS by adding $3\,\%$ to $6\,\%$ and $1\,\%$ to $2\,\%$ 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 model, but rapidly loses most of its skill over the first two forecast weeks, i.e. on the medium range, eventually leveling off thereafter on subseasonal timescales. While the CSCopt model outperforms the S2STCcal model from week three on in the central MDR, the CSCopt model 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), emphasizing the potential of climatological forecasts for subseasonal timescales.

Expanding the climatological model by including predictors generated from past data, the purely statistical approach improves the CSCopt skill at 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. 6.8a). 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. 6.8b). Considerable improvements can also be identified in

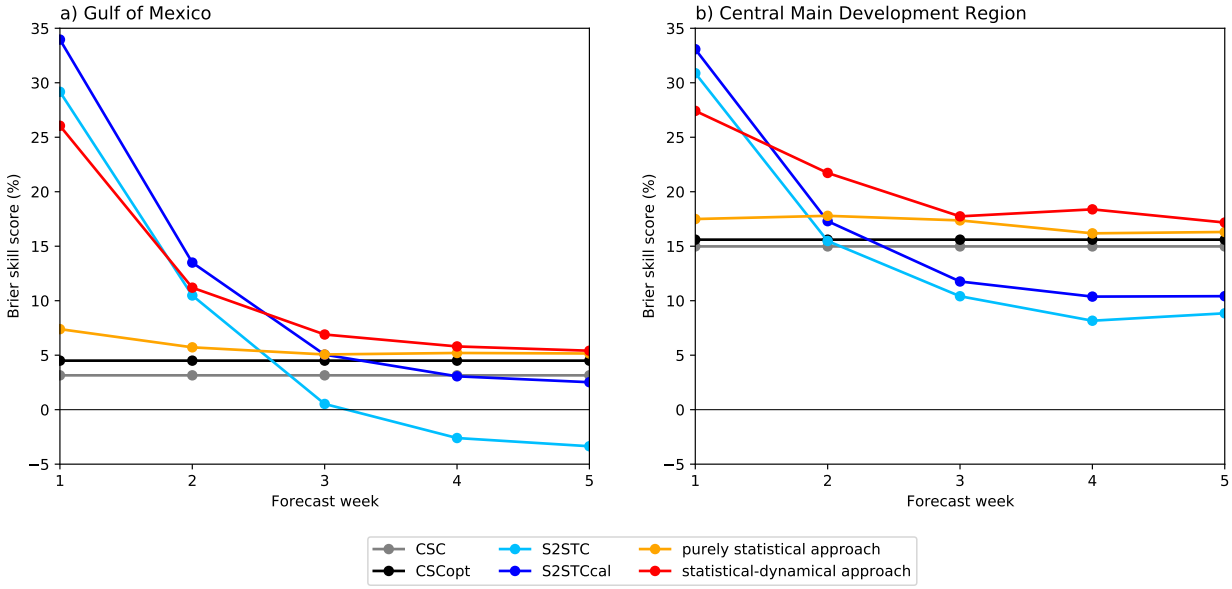

Figure 6.8: BSS (%) as a function of forecast week for the CSC (gray), CSCopt (black), S2STC (lightblue), S2STCcal (darkblue), purely statistical (orange), and statistical–dynamical (red) model, respectively, relative to the MSC model, and validated in the Gulf of Mexico (left) and central MDR (right) subregions. Reprinted from Maier-Gerber et al. (2021). © 2021, American Meteorological Society. Used with permission.

subregions defined for the Caribbean Sea, and slightly north of the MDR (not shown), suggesting that subseasonal 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 subseasonal variations in the central MDR. For the Gulf of Mexico, the improvement in skill from the purely statistical to the statistical–dynamical model is 4.5 to 6.5 (0.4 to 3.2) times greater on the medium (subseasonal) range than the improvement from the CSCopt model to the purely statistical model. For the central MDR, the ratio of improvements is 1.8 to 5.2 (0.2 to 3.8) on the medium (subseasonal) 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 equivalent skill for week one and two would be to include the S2STCcal forecasts as a predictor to the logistic regression models.

A final comparison between the above findings for potential (measured by the AUC) and actual predictive skill (measured by the BSS) corroborates the advantage of combining the two perspectives for validation. The S2STC(cal) model clearly beats the hybrid

model in terms of BSS in both subregions at week one, but lacks potential skill with regard to the AUC measure. This means that a scale-based validation (here based on the BSS) favors the dynamical model approach, whereas a rank-based validation (here based on the AUC) assesses the hybrid model approach to yield better results for week-one forecasts. On the subseasonal time scale, the dynamical model performs worse than the other model types regardless of whether potential or actual predictive skill is concerned. In the course of the ROC analysis, it has been concluded that the CSCopt model is the better choice given the high agreement of its ROC curve with those of the statistical-dynamical, the purely statistical models. This, however, cannot be confirmed in terms of actual predictive skill as the scale-based validation demonstrates the benefit of using NWP-based predictors in the hybrid model approach on subseasonal lead times.

6.3.4 Skill decomposition

As described in Section 4.4.3, the BS can be decomposed into three additive contributions, namely miscalibration (MCB), uncertainty (UNC), and discrimination (DSC), the latter entering with a negative sign. When plotting these components as stacked bars (Fig. 6.9), the models can be examined with respect to their degree of miscalibration (MCB) as well as their ability to discriminate between TC occurrence and non-occurrence (DSC). The UNC component is solely determined by the sample climatology, i.e. by the observations in the validation dataset, and hence independent of model and forecast week. The discussion will hence primarily focus on MCB and DSC components. Note that a perfect model would be able to compensate for the UNC component and a possible miscalibration by an equally large DSC component, resulting in BS=0. In the following, BS decomposition results are discussed for the central MDR, in particular to elucidate which component is crucial for the subseasonal skill of the statistical-dynamical model (i.e. reduced MCB or increased DSC). Results for model predictions in the Gulf of Mexico can be found in the appendix (Fig. A.1.6). The UNC component is slightly smaller in the central MDR than in the Gulf of Mexico (0.103 vs. 0.111), indicating that the TC occurrence prediction problem in the central MDR is the inherently easier one among the two subregions.

For the central MDR (Fig. 6.9), the first thing to note is that the DSC component is larger than the MCB component for all models and forecast weeks, which yields a BS less than the UNC component. As for the models based on the climatological seasonal cycle (Fig. 6.9a+b), the version applying the locally optimized smoothing leads to stronger miscalibration, but it overcompensates for this with an even stronger ability to discriminate between TC occurrence and non-occurrence, resulting in a better BS for the CSCopt model. The marked decline in BS of the dynamical models with increasing forecast week

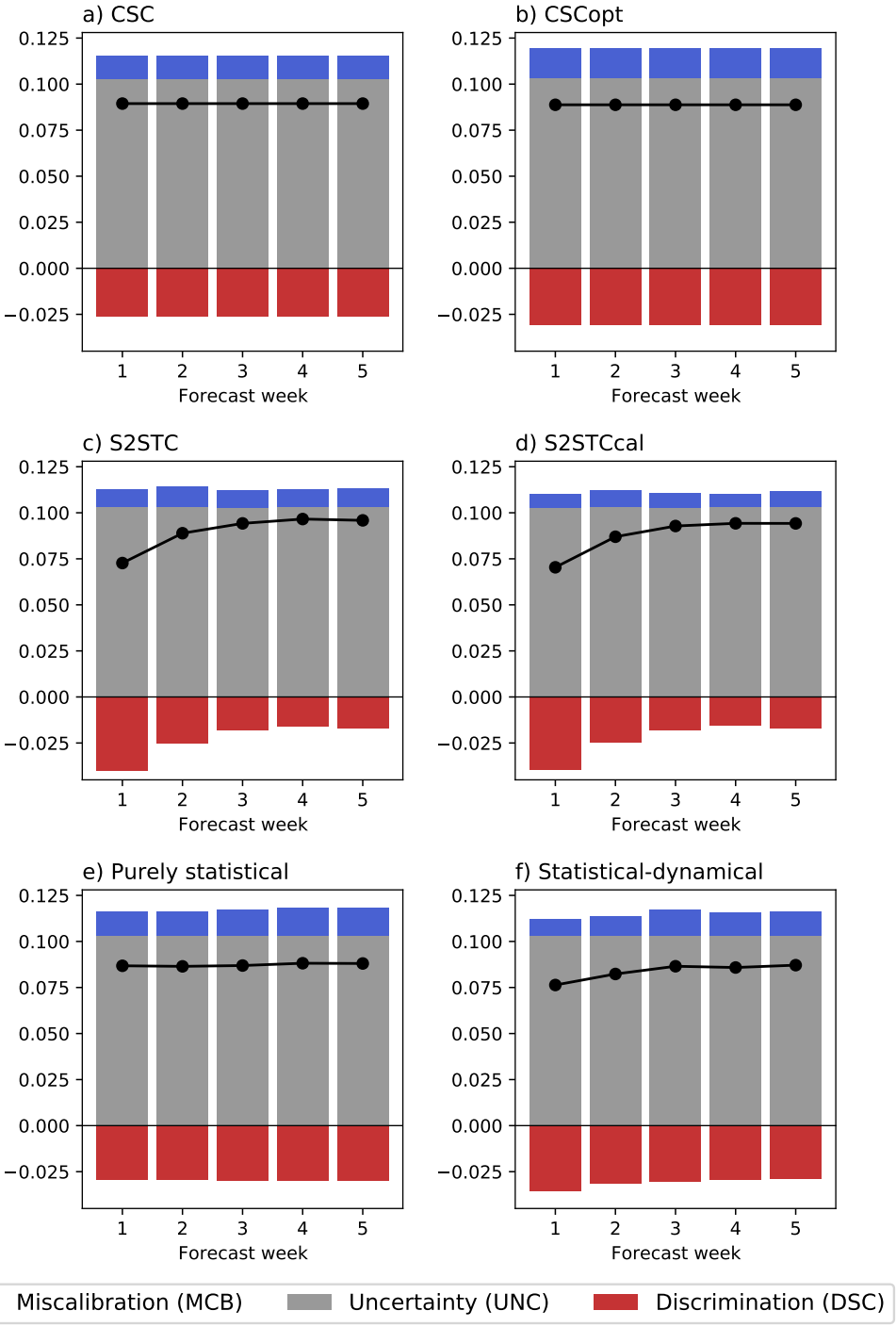

Figure 6.9: BS decomposition as a function of forecast week for the (a) CSC, (b) CSCopt, (c) S2STC, (d) S2STCcal, (e) purely statistical, and (f) statistical–dynamical model, respectively, validated in the central MDR. The BS (black line) for each week is obtained by summing up the respective uncertainty (UNC; gray), miscalibration (MCB; blue), and discrimination (DSC; red) components. Note that the DSC term has a negative sign. For more details on BS decomposition, see Section 4.4.3.

is not primarily due to an increase in miscalibration but to a decrease in discrimination (Fig. 6.9c+d). This is the reason why the IDR-based calibration cannot prevent the sharp drop in BS despite the noticeable improvement in the MCB component. Even though the S2STCcal model is the best calibrated at all leadtimes, its subseasonal predictions are less skillful compared to the purely statistical and the statistical-dynamical models, which can be attributed to its generally smaller DSC components. When compared to the CSCopt model, the decomposition patterns for the purely statistical approach are fairly similar results, whereas the dynamical model component in the hybrid approach clearly leads to a generally reduced miscalibration but also to a better discrimination on the medium range.

6.3.5 Relevance of predictor groups

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 6.10 provides insight into incremental improvements when successively including the predictor categories, summarized in Fig. 6.3, 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.

Adding the GPI predictors to the CSCopt base predictor in the Gulf of Mexico already outperforms the purely statistical approach, which chooses from the full set of past data predictors, at all lead times (Fig. 6.10a). 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 subseasonal timescale, and is almost comparable on the medium range (Fig. 6.10b). The majority of the subseasonal 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 range, another substantial fraction of the skill in both subregions results from adding information on tropical wave modes.

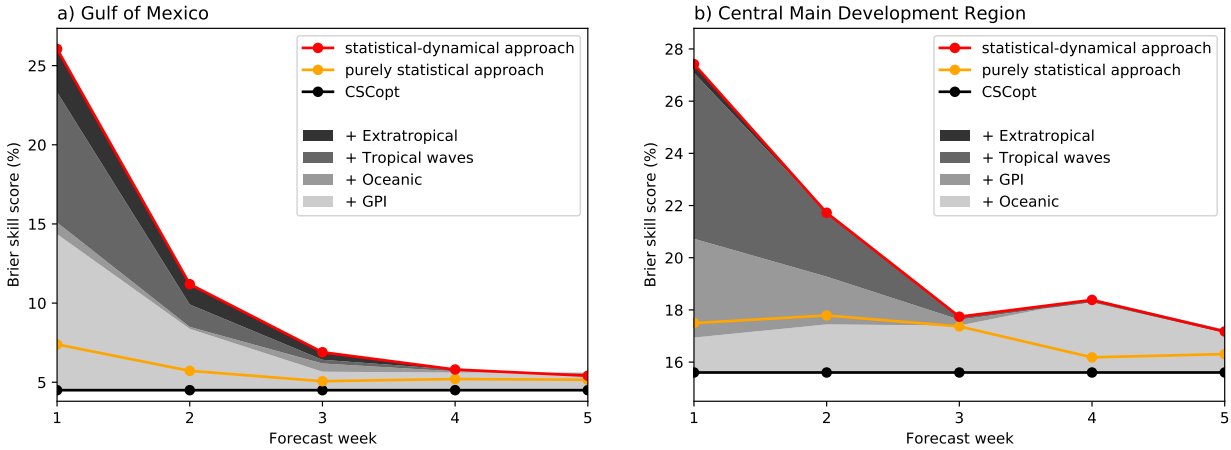

Figure 6.10: As in Fig. 6.8, but only the CSCopt, purely statistical, and statistical–dynamical models are shown in black, orange, and red lines, respectively, relative to the MSC model. For the latter, improvements in BSS, when successively including the predictor categories (see Fig. 6.3) to the sequential 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. Reprinted from Maier-Gerber et al. (2021). © 2021, American Meteorological Society. Used with permission.

6.3.6 Economic value

As described in Section 4.4.4, the choice of $p_{thres} = C/L$ results in the maximum achievable economic value only if the corresponding probabilistic forecast model is calibrated. Therefore, in accordance with the results presented in Section 6.3.1, the uncalibrated S2STC model cannot be considered in the following discussion of the envelope curves constructed for the economic values of each model. As climatological models are deemed to be calibrated by definition, the CSCopt model is not rejected for this analysis, although it has a tendency to underforecast TC occurrence for forecast probabilities exceeding the mean observed relative frequency of the respective subregion (cf. Fig. 6.6).

The overall maximum economic value V_{max} is achieved for a user with the cost-loss ratio of $C/L = \bar{y}$, and thus for 0.133 in the Gulf of Mexico and 0.118 in the central MDR, respectively, independent of the model and forecast week (Fig. 6.11). For week-1 forecasts and $C/L \gtrapprox 0.1$, the S2STCcal model has the highest economic value in both subregions, followed by the statistical-dynamical, the purely statistical, and the CSCopt models in descending order (Figs. 6.11a+b), respectively. While the purely statistical model closely resembles the curve of the CSCopt model with an extension between $C/L \gtrapprox 0.2 - 0.5$, the dynamical predictors used in the statistical-dynamical model lead to differences in V up to about 0.3 (0.2) in the Gulf of Mexico (central MDR). Because the S2STCcal envelope curve quickly drops below zero for $C/L < \bar{y}$, users located in this range will have greatest benefits when making decisions based on the statistical-

dynamical models.

For week two, the economic value associated with the S2STCcal model still slightly exceeds the one of the statistical-dynamical model in the Gulf of Mexico, but it is outperformed already in the central MDR, in particular for $C/L > 0.2$ (Figs. 6.11c+d). Another difference between the two subregions at week two is that V_{max} is almost equal for all models in the central MDR, whereas the S2STCcal and CSCopt models still differ in V_{max} by more than 0.1. With further increasing lead time, the S2STCcal forecasts lose their economic value, and eventually fall below the climatological envelope curves at week four (Figs. 6.11e+f), which is independent of the forecast week.

What stands out at week four is that no forecast model exhibits an economic value to users with a cost-loss ratio greater than 0.3 in the Gulf of Mexico and 0.5 in the central MDR, respectively. Becoming more and more congruent with the CSCopt model for greater lead time, the purely statistical and statistical-dynamical models are only useful for users with C/L located at the right-side tail of the CSCopt envelope curve, with the greatest benefit from the statistical-dynamical model in the central MDR. Figures showing results for the remaining weeks three and five can be found in the appendix (Fig. A.1.7).

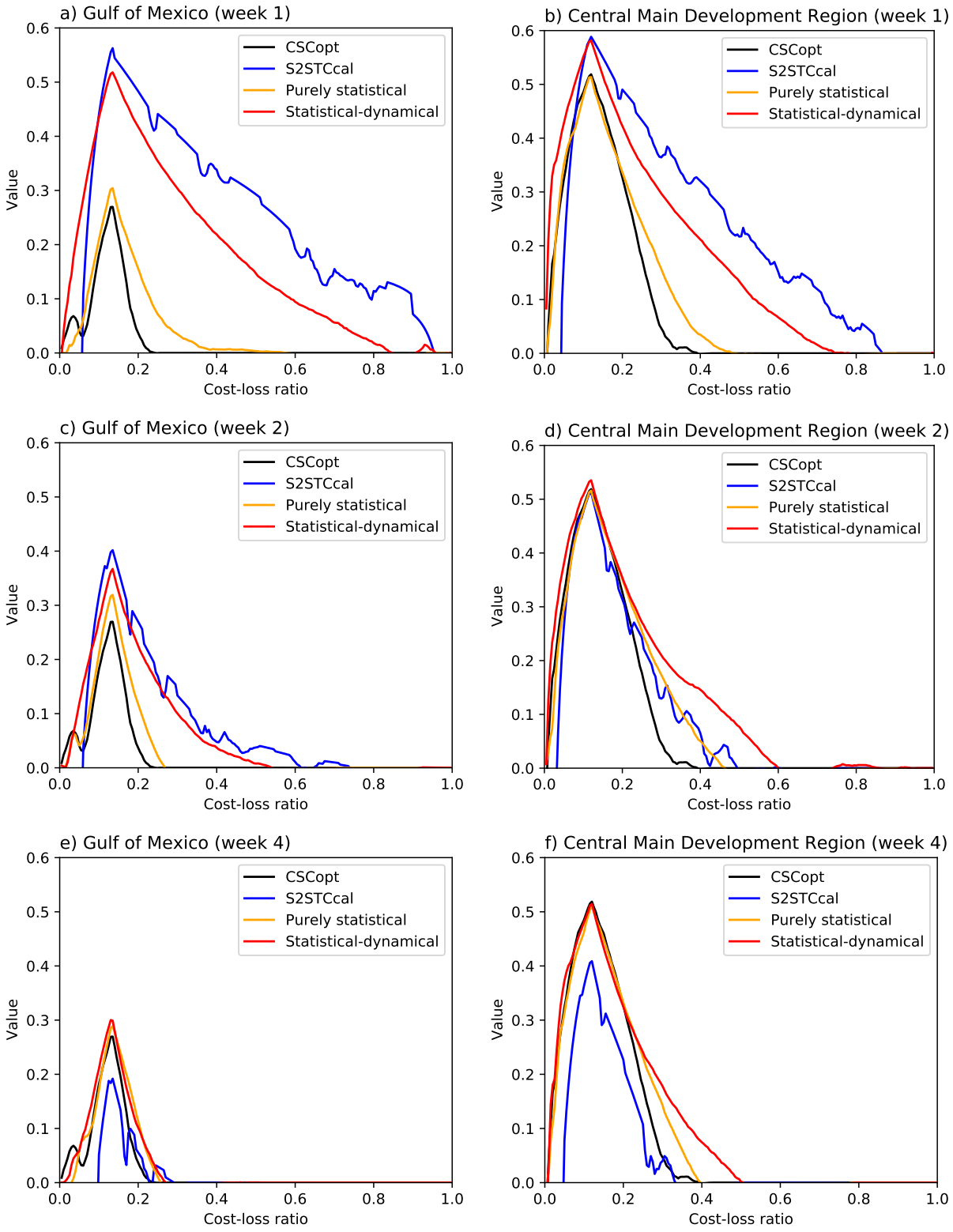

Figure 6.11: Economic value as a function of cost-loss ratio for the CSCopt (black), S2STCcal (darkblue), purely statistical (orange), and statistical–dynamical (red) models, respectively, calculated for (a+b) week-1, (c+d) week-2, and (e+f) week-4 forecasts in (a+c+e) the Gulf of Mexico, and (b+d+f) the central MDR subregions.

7. Subseasonal forecasting of accumulated cyclone energy

7.1 Benchmark models

As in Section 6.1, a set of benchmark models is created for the ACE target variable to compare the predictive performance of the newly developed models with appropriate references. With climatological models on the one hand, and NWP models on the other, two fundamentally distinct approaches are considered for that purpose, and these are described in the following in their original and optimized versions. In addition, a trivial model is deployed to provide a reference for clarifying to what extent climatological information is valuable when predicting the ACE distribution.

7.1.1 Trivial model

Probably the most trivial approach to generate a predictive probabilistic distribution for ACE is to define a model by a PDF of the form

$$\hat{p}(ACE) = \begin{cases} 1, & ACE = 0 \\ 0, & ACE > 0. \end{cases} \tag{7.1}$$

This model, hereafter referred to as *ZEROS* model, assumes certainty regarding its prediction of $ACE = 0$, and is thus equivalent to a deterministic forecast. Because its formulation does not consider any information other than the stated assumption, i.e. that no TC occurs, it serves as reference to assess how rare $ACE > 0$ actually is in the target variable, and hence how other models containing more information perform compared to this trivial approach.

7.1.2 Climatological forecasts

Under the assumption of stationarity, i.e., that the distribution of the target variable is not subject to long-term changes, climatological information can be used to define benchmark models, especially for longer lead times. The empirical CDF predicted by a climatological model is composed of past observations, and is thus independent of the target forecast week. For the present application, all climatological models are based on the 1968–2017 period of the IBTrACS dataset. In analogy to the definition of the MSC model for TC occurrence, a very simple approach to create a climatological reference model for ACE is to use all observations of the 50 North Atlantic hurricane seasons considered. Because this way the same probability distribution is predicted throughout the season, this climatological model will hence be referred to as full-season probabilistic climatology (FSPC). To allow for seasonal variations in the composition of the empirical distribution, a second climatological model combines observations from the reference period for a given day of year, and is hereafter referred to as probabilistic climatology (PC).

As was shown in the previous chapter for the CSC model, the 50-year reference period used to build the climatological models is not sufficient to obtain physically reasonable variability over the season. Because of the problem of sparse observations that are typically available for such climatological models, but with a focus on tropical rainfall, Vogel et al. (2018) and Walz et al. (2021) suggest to add past observations of $\pm x$ days relative to the day of year of interest. This inclusion of data from adjacent days is thought to make the resultant empirical climatological distribution more robust, and the approach is therefore applied as a third climatological benchmark, with the corresponding model being referred to as extended probabilistic climatology (EPC) of window length x. While EPCs with small window lengths suffer from the same problem as the PC, EPCs with larger window length increasingly loose the ability to represent seasonal variability. The optimal window length for an EPC is identified by minimizing the CRPS for each subregion separately. The best performing EPC (hereafter are referred to as *EPCopt*) is found for a window length of ± 21 days in the Gulf of Mexico, and of ± 9 days in the central MDR (Fig. 7.1). The larger window length in the Gulf of Mexico indicates that the empirical distribution composed by the EPC approach requires more data to become robust, i.e. to more closely resemble the underlying climatological distribution of ACE. In contrast, good resemblance seems to be achieved with less data necessary in the central MDR, which is likely to be due to the more pronounced seasonal cycle. These results are in qualitative agreement with those obtained for the smoothing of the CSC (see Section 6.1.1), where the optimal window length was also found to be twice as large in the Gulf of Mexico. Since all models in this study are validated in a CV mode with 20 folds

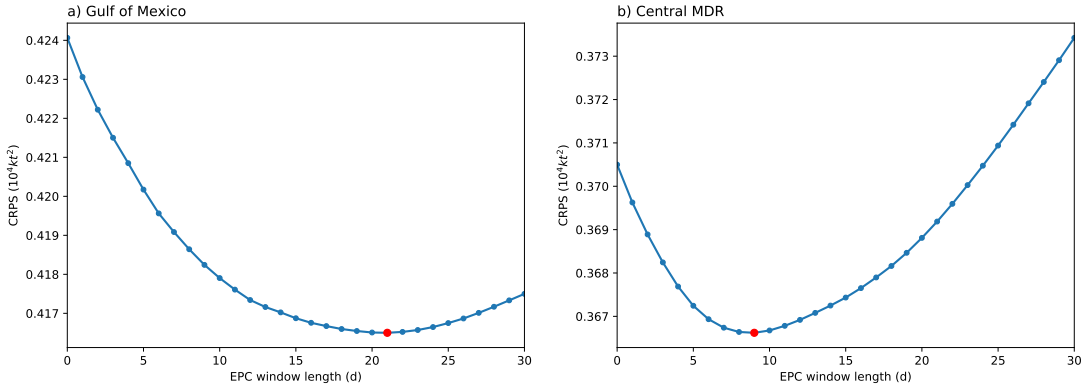

Figure 7.1: CRPS ($10^4 kt^2$) of the EPC model for a range of window lengths for (a) the Gulf of Mexico, and (b) the central MDR, respectively. The red dot indicates the minimized CRPS and thus the optimal window length determined.

(see Section 4.3), and climatological models are constructed from observations, the observations of the particular CV-fold must be excluded. This results in 20 versions of the climatological model, but differences in the composition of the probability distributions are negligible (not shown).

7.1.3 Dynamical forecasts

In addition to climatological models, S2S ensemble reforecasts are used to provide NWP-based benchmarks as a second type of reference. For these type of models, an empirical distribution of ACE is composed of the 10 perturbed forecasts and the one control forecast, i.e. from 11 members in total. As mentioned in the previous chapter, TCs identified in the S2S dataset suffer from a systematic underestimation of their intensity. While this issue could be easily fixed through a lowered threshold for tropical storm strength in case of TC occurrence prediction in Section 6.1.2, a bias correction of ACE derived from S2S forecasts requires to adujst the full TC intensity distribution. Like in Lee et al. (2018), this is achieved in the present study by applying a quantile mapping between the lifetime maximum intensity (LMI) of the 1998–2017 TCs predicted in the S2S and observed in the IBTrACS dataset, respectively. Since S2S forecasts range out to 46 days, and the longest lifetime of a North Atlantic TC recorded was 27.75 days (AOML, 2021), the forecasts should be long enough to capture a TC's LMI. Because the S2S ensemble members are exchangeable and therefore indistinguishable, the forecasted LMIs are pooled across the ensemble dimension before calculating the mapping. Figure 7.2 shows that, for any quantile, the S2S model predicts TCs to be considerably weaker than observed, with the CDF reaching 100% at $96\,\mathrm{kt}$, whereas the largest value in the observed CDF is $160\,\mathrm{kt}$.

Results of the window length optimization used for the EPC benchmark model revealed that an inclusion of additional data around a given day of year increases the pre-

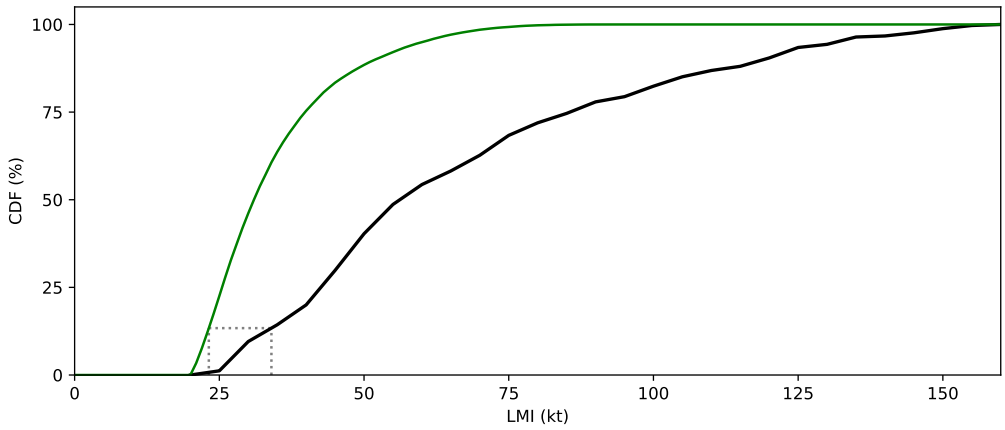

Figure 7.2: CDF (%) of the observed (black) and S2S-predicted (green) LMI, respectively. The gray dotted lines indicate the quantile-based mapping that corresponds to an observed intensity of 34 kt, i.e. the minimum intensity for a cyclone to contribute to ACE.

dictive performance, due to a more robust representation of the climatological probabilistic distribution. Since the S2S reforecasts come with an 11-member ensemble, the question arises whether this ensemble size is sufficient to resemble the underlying distribution of the target variable. To answer this question, subsets are generated consisting of n_{pert} perturbed members randomly drawn from the original 10 perturbed members without replacement, plus the one control forecast. With n_{pert} ranging from 0 (i.e. the control reforecast only) to 10 members (i.e. the full S2S reforecast ensemble), the effect of the ensemble size on the performance can be assessed (Fig. 7.3). In both subregions, the CRPS improves significantly over the control forecast with the addition of the first perturbed forecasts, but degressively thereafter. Because the S2S ensemble size is limited and the behavior for $n_{pert} > 10$ hence cannot be determined, the indication of converging CRPS curves in Figure 7.3 suggests that the 11-member ensemble should be sufficient to compose a robust empirical distribution for ACE (hereafter referred to as *S2SACE*).

A priori, it cannot be ruled out that predictions of the S2SACE model do not suffer from systematic errors. Therefore, a second version of this model is generated, hereafter referred to as *S2SACE-SPP*, statistically post-processing the original S2S ensemble forecasts. The post-processing is implemented in form of IDR, where all 11 forecast members are treated as individual covariates. Because IDR usually performs better on larger datasets, gridpoints are pooled for model training within each of the two subregions, like it is done during calibration of the predicted probabilities of the S2STC model in the previous chapter. However, the pooling strategy is dropped for the prediction mode, so that forecasts are generated by the statistical post-processing model at every gridpoint separately. To apply the full set of forecast verification tools introduced in Section 4.5, the predictive distribution, which is given in form of a piecewise constant CDF, is repre-

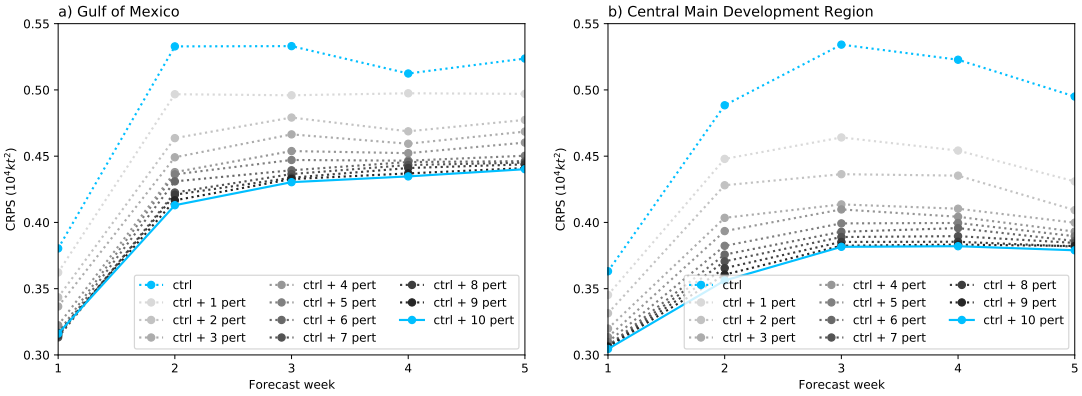

Figure 7.3: CRPS ($10^4 kt^2$) for different sets of randomly drawn S2S sub-ensemble reforecasts (see text for details).

sented by a sample of size 1000 for pragmatic reasons. The sampling units of the unkown PDF are generated through inverse transform sampling, a method in which random draws from a standard uniform distribution are used as input for the inverse CDF of the predictive distribution.

7.2 Statistical model development

7.2.1 Two-part modelling approach

As pointed out in Section 4.1.1, the key challenge for statistical models in predicting the probability distribution of the target variable ACE is to accurately forecast the large point mass at zero along with the quickly dropping distribution for $ACE > 0$. Due to the discontinuity at $ACE = 0$ and the gap resulting from the required minimum TC intensity ($\geq 34\,\mathrm{kt}$), we split the modelling of the distribution of ACE into two parts. Since logistic regression models have been developed in the previous chapter already, it stands to reason that they are reused for the first part to predict the probability for TC occurrence $\hat{p}_{TCocc}$. As described in detail in the Section 7.2.2, the second part is then concerned with modelling the distribution for $ACE > 0$ using a truncated regression – an approach often used for censored target variables (e.g., Cragg, 1971; Tobin, 1958).

Similar to the S2SACE-SPP model, the predictive distribution of the two-part model is represented through a finite sample of size 1000 for validation purposes. This sample size is sufficient for representation, since deviations in validation results are negligible for larger sample sizes, and interpretations remain unchanged. Technically, the sample consists of $1000 \times (1 - \lfloor \hat{p}_{TCocc} \rceil)$ sampling units for $ACE = 0$, and $1000 \times \lfloor \hat{p}_{TCocc} \rceil$ samples for $ACE > 0$. The latter are obtained through inverse transform sampling, as is done in case of the S2SACE-SPP model (cf. Section 7.1.3).

7.2.2 Truncated logistic distribution regression

As motivated in Section 2.3.1, carefully selecting the type of distribution is a key element for parametric modelling. Although ACE is used in many studies, rarely an attempt is made to model its distribution. Villarini and Vecchi (2012) though test gamma, Gumbel, lognormal, and Weibull distributions (Krishnamoorthy, 2006) when developing a statistical-dynamical model for seasonally aggregated basinwide ACE, and find the gamma distribution to perform best. However, our gridded framework with a $7.5°$-radius for TC evaluation and the aggregation over individual forecast weeks yields a distribution for ACE that certainly differs from the one modelled by Villarini and Vecchi (2012). While seasonal ACE usually takes values greater than zero, as it is climatologically very likely that at least some TCs occur during the hurricane season, the probability distribution for weekly ACE is characterized by a large point mass at zero. But since we do not consider $ACE = 0$ for the second part of the model, weekly ACE is rendered into a positive target variable and we hence fit a logistic distribution left-truncated at zero. The logistic nature is chosen as the definition of ACE is tightly coupled to wind speed, and the truncated logistic distribution has turned out to be advantageous to other predictive distributions when post-processing ensemble forecasts for wind speed (Scheuerer and Möller, 2015). But because wind speed is used squared in the definition of ACE, the distribution of ACE is more skewed and has a higher kurtosis compared to the distribution of wind speed. The fact that ACE sums over multiple instances, however, somewhat mitigates these deviations, so that it is reasonable to choose the truncated logistic distribution for modelling weekly ACE as well. A truncated normal distribution has also been tested but was no longer pursued due to frequently raised convergence issues occurring in the boosting process, which is described in Section 7.2.3.

The PDF of the logistic distribution is defined as

$$\hat{p}(y, \mu, \sigma) = \frac{e^{-(y-\mu)/\sigma}}{\sigma \left(1 + e^{-(y-\mu)/\sigma}\right)^2}, \quad y \in \mathbb{R}, \tag{7.2}$$

and the corresponding CDF as

$$\hat{P}(y, \mu, \sigma) = \left(1 + e^{-(y-\mu)/\sigma}\right)^{-1}, \tag{7.3}$$

where μ and σ are the location and scale parameters, respectively, thus constituting a two-parametric model. Following the guideline of keeping the model as simple as possible (cf. Section 2.3.1) and in view of the lower predictability on subseasonal time scales, accurate modelling of these two parameters, representing the first and second statistical moments, is considered sufficient.

When ACE is truncated at zero, the corresponding left-truncated PDF is obtained by modifying Eq. (7.2) as follows

$$\hat{p}_{trunc}(y, \mu, \sigma) = \frac{\hat{p}(y, \mu, \sigma)}{1 - \hat{P}(0, \mu, \sigma)}, \quad y > 0, \tag{7.4}$$

where y is the ACE. As with the development of the logistic regression models for TC occurrence in Chapter 6, and thus the first part of the two-part model, the difference between the statistical-dynamical and purely statistical approach in the second part also lies in the predictor set that is used to define the link functions. Using the predictors developed in Chapter 5, the location parameter μ is modelled as

$$\mu = \boldsymbol{x}^T \boldsymbol{\alpha} = \alpha_0 + x_1\alpha_1 + \cdots + x_P\alpha_P, \tag{7.5}$$

where $\boldsymbol{x} = (1, x_1, \ldots, x_P)^T$ contains an intercept term and the values x_i, $i = 1, \ldots, P$, of the P predictors, and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_P)^T$ is the vector including the intercept α_0 and the regression coefficients α_i, $i = 1, \ldots, P$, of the predictors.

Since there is no reason to assume the uncertainty in the ACE distribution modeled by the scale parameter σ to be constant, a heteroscedastic model is defined, i.e., σ also depends on various predictors. In ensemble model output statistics (EMOS; Gneiting et al., 2005), the usual approach is to include the ensemble standard deviation of the forecast variable as a predictor. We want to be less restrictive a priori in terms of potentially relevant predictor variables and define a scale model that uses the full set of developed predictors.

$$\log(\sigma) = \boldsymbol{x}^T \boldsymbol{\gamma} = \gamma_0 + x_1\gamma_1 + \cdots + x_P\gamma_P, \tag{7.6}$$

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_P)^T$ is the vector including the intercept γ_0 and the regression coefficients γ_i, $i = 1, \ldots, P$, of the predictors. Note that a logarithm transformation is used in Eq. (7.6) to ensure that $\sigma > 0$. While maximum likelihood estimation was applied to fit the logistic regression model for TC occurrence in Chapter 6, the CRPS is used to formulate the loss function when estimating the parameters defined in Eqs. (7.5) and (7.6). Maximum likelihood estimation was also tested but resulted in either comparable or less skillful models than those for which the CRPS was minimized.

7.2.3 Solver-integrated predictor selection

Although the optimization of the loss function could be performed with a variety of solvers, the nonhomogeneous gradient boosting algorithm proposed by Messner et al. (2017) is deployed here for the following reason. Equations (7.5) and (7.6) are formulated such that all developed predictors are included in the model for the location and

scale parameter, respectively. To avoid including too many predictors, which is prone to overfitting, a strategy is needed to automatically select only relevant predictors. While for the statistical TC occurrence models in Chapter 6 relevant predictors are determined by using sequential predictor selection, we refrain from applying this time-consuming optimization scheme when selecting predictors for the truncated logistic distribution model for ACE. Using the nonhomogeneous gradient boosting algorithm has the advantage that only the most relevant predictors are included in the model during the iterative optimization of the loss function, although the entire predictor pool is provided to the algorithm.

Initially, regression coefficients in $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are set to zero, and then only the coefficient of the currently most relevant predictor is adjusted in each iteration. The relevance of each predictor is determined through calculating the correlation of its value with the negative partial derivative of the loss function with respect to either μ or σ, depending on whether the predictor is included in Eqs. (7.5) or (7.6). For each of these equations, the predictor with the highest correlation is considered to be most relevant. The corresponding coefficient is then modified by adding to the current value the correlation coefficient multiplied by a factor for the step size ν. Bühlmann and Hothorn (2007) suggest to use $\nu = 0.1$ for the step size, but $\nu = 0.05$ has proven to be more suitable for the current application, as a reduced boosting increment makes the convergence more stable. Note that all predictors and the target variable are standardized beforehand based on the training data, since only then the gradient can be linearly approximated by using the correlation coefficient. In the final step of each iteration, it is examined whether either the modification of the best predictor from the location model or the scale model yields a lower CRPS, and thus only the regression coefficient of the overall best-performing predictor is actually updated.

This process is repeated a predefined maximum number of iterations $mstop_{max}$. In contrast to Bühlmann and Hothorn (2007), we perform the boosting with a three times larger $mstop_{max} = 300$ for the statistical-dynamical model to compensate for the halved step size ν, and thus for the reduced boosting increments. Because the number of predictors in the predictor pool of the purely statistical model is only half as large as for the statistical-dynamical model, $mstop_{max} = 150$ is used for the purely statistical model to achieve a roughly comparable degree of maximum boosting. Once the boosting algorithm is stopped after $mstop_{max}$ iterations, we determine the optimal number of iterations $mstop_{opt}$ by the iteration with the lowest AIC. Ideally, it represents a good balance between too few predictors selected (resulting in a model lacking skill by running too few iterations), and too many predictors selected (resulting in overfitting by running too many iterations). Using the AIC as the stopping criterion complements the CRPS-based loss function, as this adds the aspects of model performance and prevention of overfitting to the aspects of improving calibration and sharpness during model selection.

The remainder of this subsection presents and discusses the resulting standardized coefficients averaged over the 20-folds of the CV. Although the full predictor pools are provided to the location model, for both subregions and model approaches, none of them is selected to enter Eq. (7.5) with a non-zero coefficient. Even the probability for TC occurrence is not selected for the location model when provided as an additional predictor (not shown), which highlights the challenge of predicting weekly ACE. The location model hence degenerates to the intercept, $\mu = \alpha_0$, where α_0 takes the value -39.89 (-41.41) in the Gulf of Mexico (central MDR), not varying with forecast week or model approach. The location model being independent of the predictors developed together with the fact that several predictors are selected as relevant for the scale model (as will be discussed in the following) proves that the prior assumption of heteroscedasticity is not just well founded but essential for appropriate modelling of the probability distribution of ACE.

While an l_2-regularization term is used for the logistic regression model for TC occurrence (see Section 6.2.1), the nonhomogeneous gradient boosting algorithm applied in this chapter more closely resembles the behaviour of an l_1-regularization (Hastie et al., 2009). In the case of multicollinearity between predictors, this results in one of every two strongly correlated predictors being chosen in preference to the other. For this reason, it is important to point out that the relevance of the selected predictors discussed below cannot be understood as conclusively causal, but only as their importance for optimizing the statistical model developed. A more in-depth analysis would be necessary to reveal causal relationships, but this is beyond the scope of this study. On the one hand, predictor coefficients close to zero may indicate a lack of relevance, but on the other, it is also possible that some coefficients of the 20 folds have opposite sign, thus cancelling out when their regression coefficients are averaged across folds. The discussion will hence focus on predictors, for which the absolute amount of the standardized coefficients exceeds 0.1, and only mention their averaged sign if negative.

Figure 7.4 shows the results of the solver-integrated predictor selection for the scale model of the statistical-dynamical approach. The intercept γ_0 (not shown) for week 1, 2, 3, 4, and 5 is $0.77, 0.96, 0.99, 0.99,$ and 0.97 in the Gulf of Mexico, whereas it is $0.57, 0.59, 0.47, 0.48,$ and 0.50 in the central MDR, respectively. Except for the mean Kelvin wave predictor in the Gulf of Mexico, every predictor has been selected at least once in any of the five forecast weeks considered.

In the Gulf of Mexico (Fig. 7.4a), the mean MDR SST is found to be the most relevant predictor throughout all forecast weeks, with standardized coefficients ≥ 0.2 at week 1 and ≥ 0.3 beyond. At week 3 and 4, the corresponding predictor of the ensemble standard deviation also contributes more strongly. In terms of the tropical wave predictors, coefficients are mostly positive. Greater relevance appears to be for the MRG/TD wave

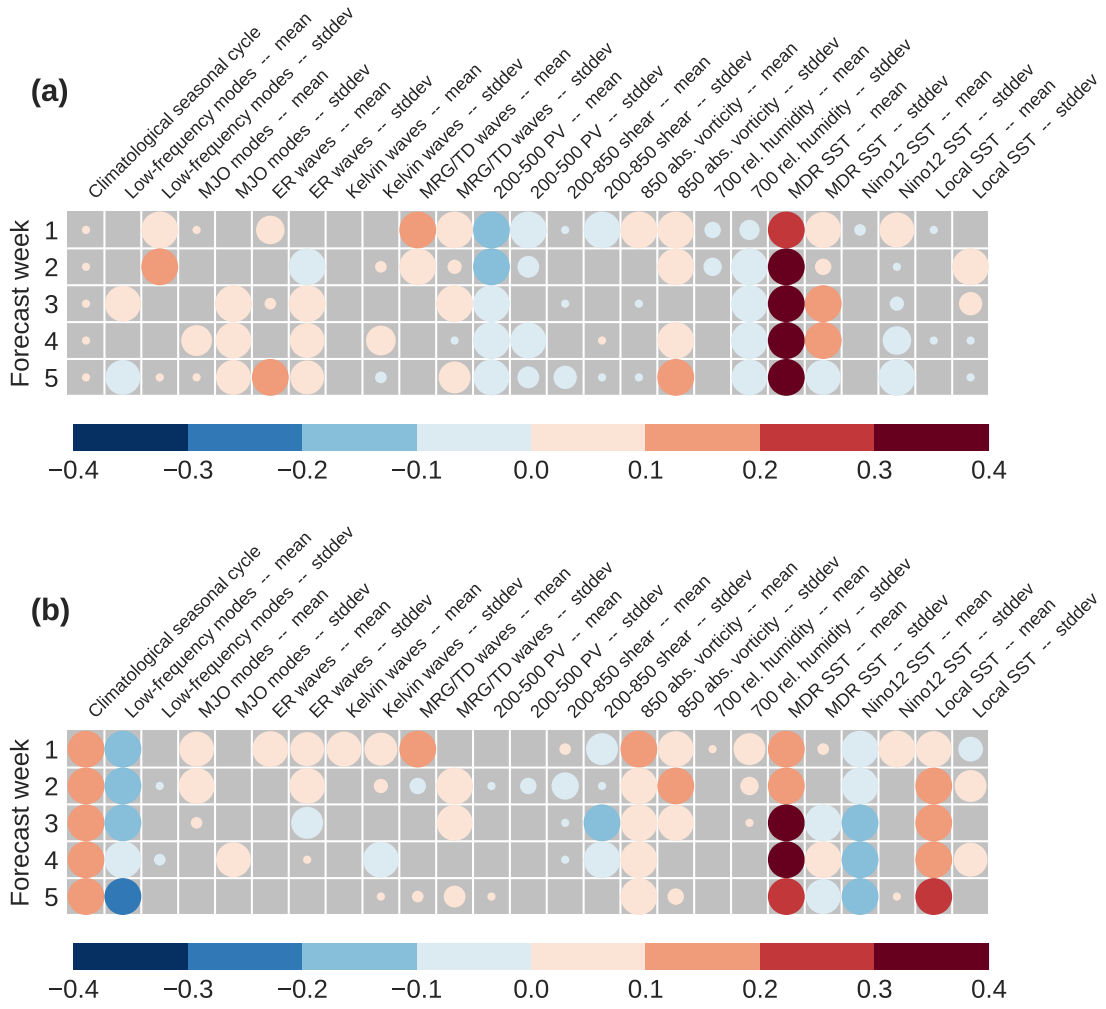

Figure 7.4: Predictor relevance in the scale model determined by the solver-integrated selection for the statistical-dynamical approach for week one to five in (a) the Gulf of Mexico, and (b) the central MDR, respectively. Colors indicate the standardized regression coefficient of the selected predictor averaged over the 20 folds of the cross-validation. The area of each circle is proportional to the number of folds having selected the predictor, with the diameter spanning the width of gray square if all folds have it. If no circle is plotted, the predictor was not chosen in any of the folds. Note that the intercept coefficient is not shown but stated in the text.

mean predictor in week 1, probably due to the imprint of the TCs forecast by the S2S model, which themselves project onto this wave type. However, beside standard deviation of low-frequency modes at week 2 and mean ER waves at week 5, no prominent relevance can be attributed to tropical waves in this subregion otherwise. Furthermore, the standard deviation of 850-hPa absolute vorticity contributes with higher coefficients in week 5, and the mean 200-500-hPa PV with negative coefficients on the medium range.

In the central MDR (Fig. 7.4b), coefficients for the mean MDR SST predictor are

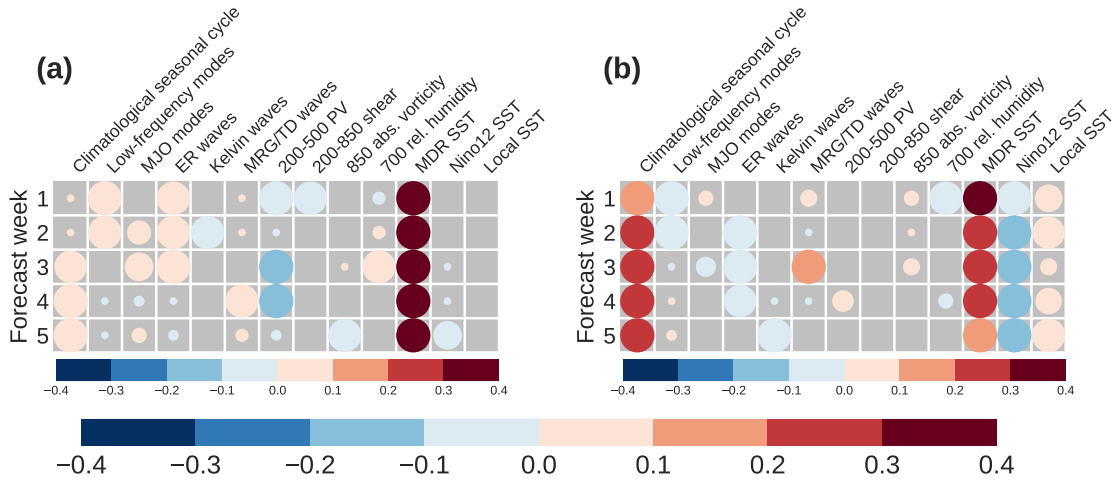

Figure 7.5: As in Fig. 7.4, but for the purely statistical approach.

also strikingly high, but with greater relevance only after week 2 as opposed to the Gulf of Mexico. The relevance of the mean MDR SST predictor on the subseasonal time scale is accompanied by the mean local SST predictor, as well as negative coefficients for the mean Nino12 SST predictor. Thus, the oceanic predictor group turns out to be crucial for accurately predicting the uncertainty of the ACE distribution in the central MDR. While the climatological seasonal cycle is selected from all 20 folds in each week, it is considered relevant by only one fold in each week in the Gulf of Mexico. This suggests that the more pronounced climatological seasonal cycle of TC occurrence in the central MDR not only provides a better base predictor for TC occurrence compared to the Gulf of Mexico (shown by the larger difference in BSS between the MSC and CSCopt models in Fig. 6.8), but that the associated predictive potential translates to the prediction of ACE as an aggregate variable. Furthermore, the mean low-frequency modes predictor (with negative coefficients) is selected in all weeks except week four. The mean MRG/TD wave at week 1 is the only relevant tropical wave predictor. As for the GPI predictors, standard deviation of 200-850-hPa vertical wind shear at week 3, as well as 850-hPa absolute vorticity (mean & stddev) at week 1 and 2, respectively, are associated with more relevance for the scale parameter.

The determined relevance of predictors included in the scale model for the purely statistical approach is shown in Fig. 7.5. The intercept γ_0 (not shown) for week 1, 2, 3, 4, and 5, is $0.99, 1.02, 1.04, 1.06,$ and 1.05 in the Gulf of Mexico, whereas it is $0.68, 0.69, 0.74, 0.73,$ and 0.74 in the central MDR, respectively. In the Gulf of Mexico, the local SST predictor is the only one that is assessed to be consistently irrelevant, whereas such non-consideration occurs only for the 200-850-hPa vertical wind shear in the central MDR. Like in the statistical-dynamical approach, the most outstanding predictor relevance in the Gulf of Mexico, with standardized coefficients (≥ 0.3) throughout

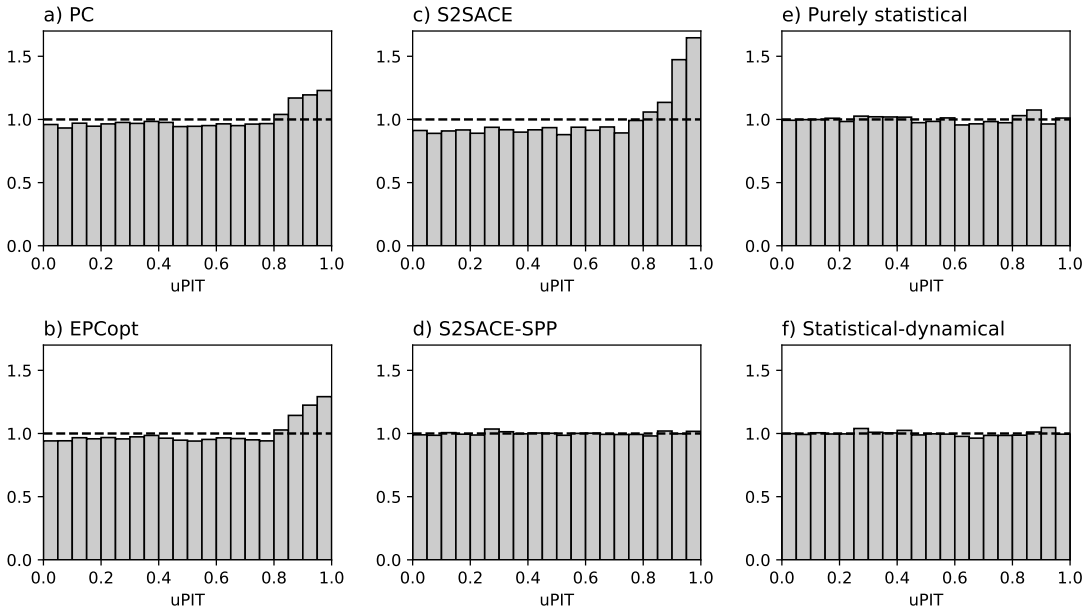

Figure 7.6: uPIT histograms for (a) PC, (b) EPCopt, (c) S2SACE, (d) S2SACE-SPP, (e) purely statistical, and (f) statistical-dynamical forecasts at week four in the Gulf of Mexico. The dashed line in each panel highlights a standard uniform distribution and thus serves as visual reference to assess miscalibration.

all forecast weeks, can be attributed to the MDR SST. The second largest (negative) coefficients are for 200-500-hPa PV at week 3 and 4, but there are no salient contributions beyond that. Although the MDR SST plays a special role in the central MDR as well, its relevance decreases with increasing forecast time. Another predictor in this subregion that is of great importance for all forecast weeks is again the climatological seasonal cycle. Compared to the statistical-dynamical approach, the Nino12 SST predictor retains its attributed relevance, but the local SST receives lower coefficients. Apart from these consistently relevant predictors, a higher coefficient is determined at week 3 for the MRG/TD wave.

7.3 Model comparison

7.3.1 Calibration

All models forecasting the distribution of ACE are first assessed in terms of their calibration before the validation is concerned with their predictive skill. As described in Section 4.5.1, this is carried out by calculating PIT values and analyzing their distribution on the unit interval. Figure 7.6 depicts uPIT histograms for week-four forecasts in the Gulf of Mexico. Week four has been chosen to examine calibration on a subseasonal lead time, but results are qualitatively consistent for the other forecast weeks considered

(not shown). Comparing with the shape for a perfectly calibrated model, which is indicated by the dashed line, the distributions of the displayed histograms (Fig. 7.6a-c) reveal an underforecasting by the probabilistic climatologies (PC and EPCopt) and the non-post-processed S2S forecasts (S2SACE). The reason that strong deviations only occur for bins $\gtrsim 0.8$, and not over the entire unit interval (as in Fig. 4.6) is that there is generally a high number of forecast instances where the observation and a large portion of the sample units are tied, in which case the PIT values are drawn randomly (see Section 4.5.1 for details). While the miscalibration of the S2SACE model is not surprising, since NWP predictions are frequently uncalibrated and hence must be corrected by statistical post-processing, climatological models should be calibrated by definition, as they are obtained from past observations of the target variable. As was speculated for the underforecasting bias of the CSCopt model when predicting TC occurrence (cf. reliability diagrams in Fig. 6.6), the miscalibration can probably be explained by a positive climatological trend between the extension (1968-1997) and the validation (1998-2017) period. Applying IDR-based statistical post-processing to the uncalibrated S2SACE model indeed yields a well-calibrated S2SACE-SPP model (Fig. 7.6c+d). The uPIT histograms for the purely statistical and statistical-dynamical approach (Fig. 7.6e+f) indicate that the two-part model developed produces well-calibrated predictive distributions for ACE, due to the fact that the optimization of the loss function is based on the CRPS. The distributions of the uPIT histograms for the central MDR (Fig. 7.7) are overall consistent with the results for the Gulf of Mexico, and so is their interpretation. The most notable difference is in the degree of miscalibration, as the two probabilistic climatologies and the S2SACE model exhibit smaller deviations in the rightmost bins of the uPIT histograms (Fig. 7.7a-c) compared to the distributions for the Gulf of Mexico.

7.3.2 Potential predictive skill

The predictive skill of the competing models shall first be assessed in terms of potential skill using UROC curves. Although the UROC curve concept generalizes from the binary to the real-valued case, it is only applicable to single values but not to full distributions as predicted for ACE in this study. UROC curves are hence constructed for the mean, i.e. the first statistical moment, of the forecast distribution in the following, to realize a first-order validation, keeping in mind that statements regarding potential predictive skill cannot be conclusive. Figure 7.8 provides UROC curves and corresponding CPA values for four distinct models at week one and four in the two validation subregions, respectively. Results for the remaining forecast weeks can be found in the appendix (Fig. A.2.3). Note that results for the S2SACE-SPP model are not presented and discussed separately as uROC curves, and thus CPA, are invariant under strictly monotonic transformation

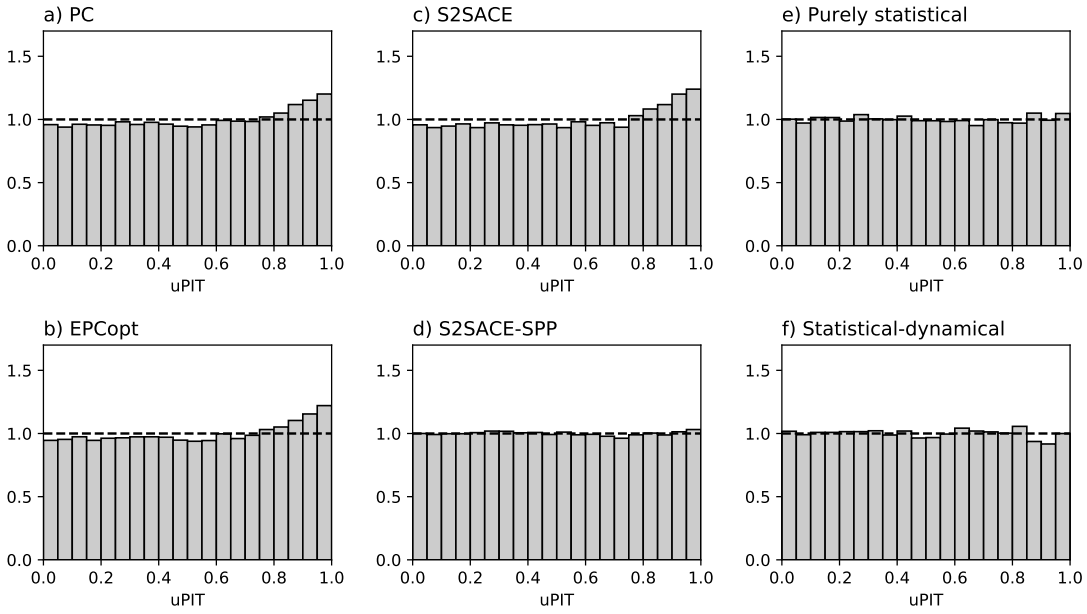

Figure 7.7: As in Fig. 7.6, but for the central MDR.

(Gneiting and Walz, 2019), which consequently applies to IDR-based post-processing (Henzi et al., 2021). Thus, all results discussed for the S2SACE model are likewise valid for the S2SACE-SPP model.

All models displayed have skill clearly better than random prediction, as indicated by the fact that all the curves are above the diagonal and therefore CPA values well are above 0.5. At forecast week 1 (Fig. 7.8a+b), the statistical-dynamical model reaches the highest CPA value among all models in the Gulf of Mexico (0.85) and the central MDR (0.87). While the S2SACE model is only 0.01 and 0.02 worse, respectively, there is a substantial difference between the subregions in terms of how much worse the EPCopt model performs. The difference in CPA values between the statistical-dynamical and the EPCopt model is 0.16 in the Gulf of Mexico and 0.03 in the central MDR, highlighting a greater potential in the Gulf of Mexico at shorter lead times for improving forecast skill over a climatological model by exploiting S2S-based predictors. Despite minor differences in CPA values, the statistical-dynamical and the S2SACE model differ in the shape of their UROC curves, in that the S2SACE model performs better for false positive rates $\lesssim 0.2$ whereas the statistical-dynamical is better for $\gtrsim 0.2$. For the most part, the UROC curve of the purely statistical model is nearly congruent with the curve of the EPCopt model, except for some deviations at very low and high false positive rates. Thus, the hybrid approach clearly has greater potential predictive skill than the purely statistical approach on shorter time scales.

Because the EPCopt model, and thus its CPA, is independent of the forecast week, it serves as a helpful visual reference for the assessment of how much the UROC curves of

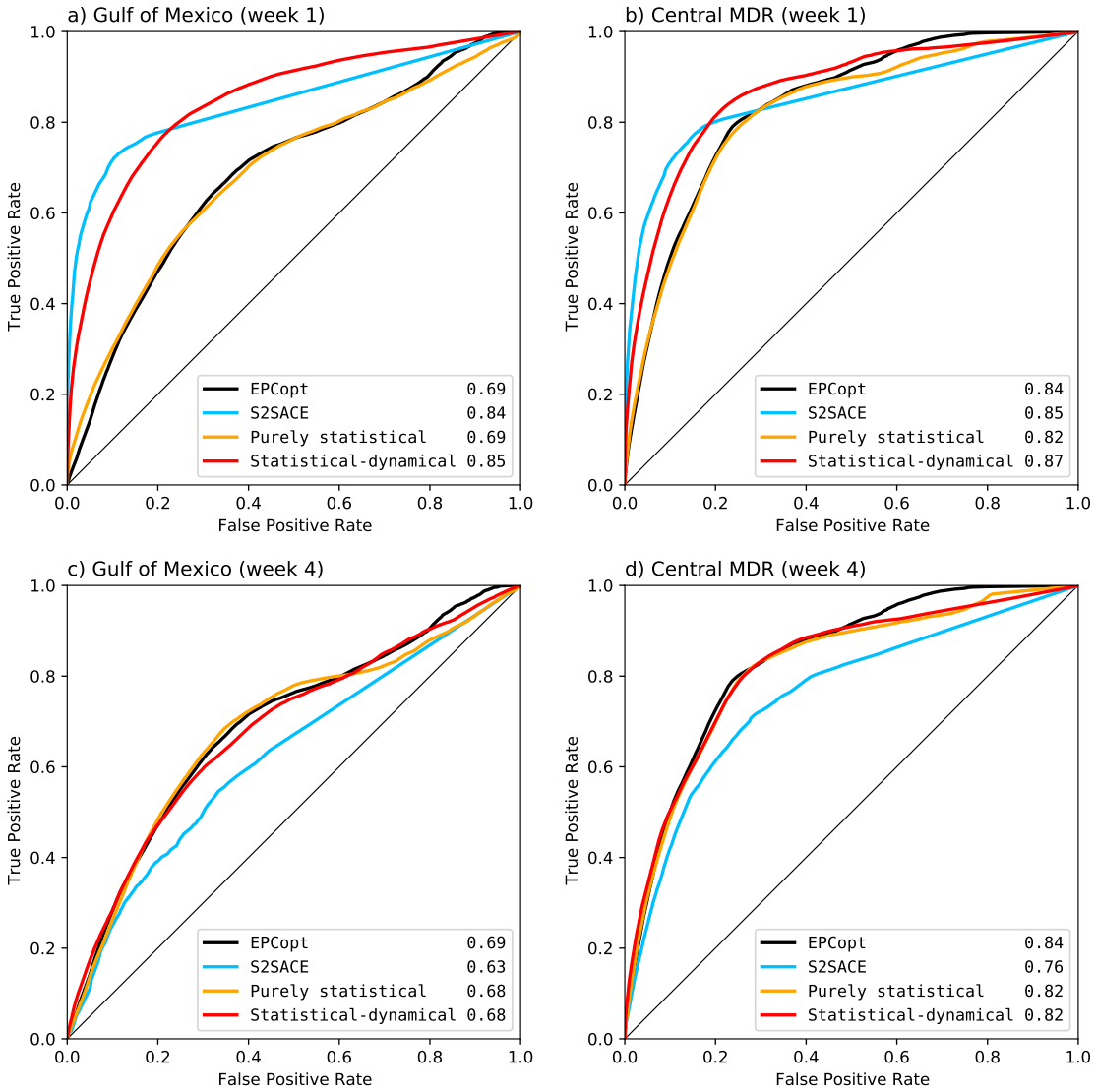

Figure 7.8: UROC curves for the forecast mean of the EPCopt (black), the S2SACE (light-blue), the purely statistical (orange), and the statistical–dynamical (red) model, respectively, at week one (a+b) and week four (c+d) in the Gulf of Mexico (a+c) and central MDR (b+d) subregion. The numbers in the legend of each panel show the model-specific coefficients of predictive ability (CPA) averaged over the 20 folds.

the other three models presented degenerate from week one to four. At week four (Fig. 7.8c+d), the curve for the S2SACE model drops well below the EPCopt curve, reaching a CPA of only 0.63 in the Gulf of Mexico and 0.76 in the central MDR. Although the potential skill of the statistical-dynamical and the purely statistical models is also noticeably reduced in week four, their UROC curves only fall back to the EPCopt model, with the climatological model still performing slightly better for high false positive rates. Both statistical models only reach a CPA of 0.68 in the Gulf of Mexico and 0.82 in central MDR, and are slightly worse (by $1-2\%$) compared to the EPCopt model. Comparing

the week-four CPA values across the two subregions suggests that there is greater subseasonal potential predictive skill for predicting the forecast mean of the ACE distribution in the central MDR than in the Gulf of Mexico.

7.3.3 Actual predictive skill

As opposed to UROC curves, which can only be applied to a single value (such as the forecast mean), the CRPS assesses the predictive performance of the full distribution, while taking into account a model's calibration and sharpness. Figure 7.9 shows a comparison of the CRPSS with respect to the FSPC model as a function of forecast week for the different model types validated in the Gulf of Mexico and central MDR subregion, respectively. How difficult it is to forecast $ACE > 0$ in general can be seen by the ZEROS model, which assumes a CDF in form of a step function (equivalent to a deterministic forecast) at 0 as a trivial approach, only performing $5 - 6\%$ worse than the FSPC reference model. It shows that TC occurrence, and thus $ACE > 0$, is so rare that predicting $ACE = 0$ with certainty is a reasonable trivial starting point. When climatological information is available, composing the predictive distribution with respect to the seasonal cycle (as in the PC and EPCopt models) leads to better skill than based on past observations of the full season (as in the FSPC model). Although these improvements are found for both subregions, with a greater benefit appearing in the central MDR than in the Gulf of Mexico (plus $7-8\%$ vs. $1-3\%$) due to the more distinct seasonal cycle, they are overall weaker compared to the improvements found when considering the seasonal cycle for TC occurrence prediction (cf. Fig. 6.8). This suggests that modelling the seasonal cycle is more relevant for the aspect of TC occurrence than for the additional intensity-related aspect inherent in ACE. As is speculated for their equivalent TC occurrence models (CSC and CSCopt in Chapter 6), the reason that the relative improvement between the PC and the EPCopt models is considerably larger in the Gulf of Mexico is likely due to the larger window length found to be optimal (21 vs 9 days) when composing the EPC for this subregion, resulting in a greater relative skill improvements.

Concerning the S2SACE model, the IDR-based statistical post-processing enhances the actual predictive skill by adding $2.0\,\%$ to $4.5\,\%$ in the Gulf of Mexico and $0.6\,\%$ to $3.1\,\%$ in the central MDR. As one would expect in terms of the performance of a dynamical model, the two versions of the S2S model clearly outperform all climatological models at week one, with the S2SACE-SPP model reaching a CRPSS of 31% and 25% in the Gulf of Mexico and central MDR, respectively. However, their skill quickly drops within the medium range forecast horizon, and the best performing climatological (EPCopt) model already exhibits better skill than the statistically post-processed NWP (S2SACE) model beyond forecast week two. This transition in the lead after about two

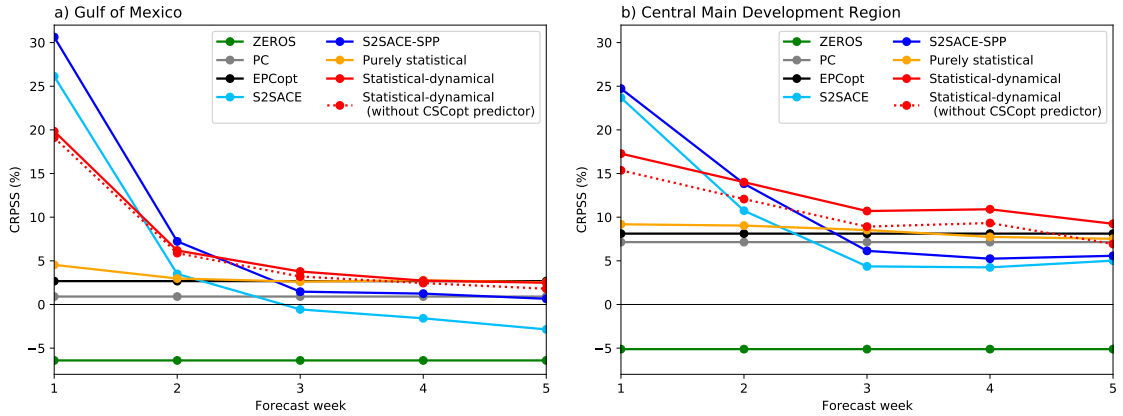

Figure 7.9: CRPSS (%) as a function of forecast week for the ZEROS (green), PC (gray), EPCopt (black), S2SACE (lightblue), S2SACE-SPP (darkblue), and purely statistical (orange) models, as well as for the statistical–dynamical model with (red solid) and without the CSCopt predictor (red dotted), relative to the FSPC model, and validated in (a) the Gulf of Mexico, and (b) the central MDR subregion, respectively.

weeks is consistent with, and hence likely determined by, the regime change that was described for TC occurrence prediction in Chapter 6.

The two-part model used in the purely statistical mode, with predictors generated from the ERA5 dataset as described in Chapter 5, overall barely improves skill over the EPCopt model. In the Gulf of Mexico, it yields a maximum of additional 1.8% at week one, but from week two on, there are no considerable deviations from the CRPSS values of the EPCopt model anymore. Although a maximum of only 1.1% is added in the central MDR, this approach performs slightly better than the EPCopt model up to week three and drops below only afterwards. While the purely statistical model proposed by Leroy and Wheeler (2008) and Henderson and Maloney (2013) in the context of TC occurrence prediction provided improvements at all five forecast weeks considered, its application to predict the distribution of ACE proves to be of little use. It should be noted, however, that unlike the previous chapter in which the CSCopt model was used as a base predictor, the EPCopt model does not enter the two-part model as a predictor because itself has the form of a full distribution.

If predictors are derived from S2S forecasts, thus constituting the statistical-dynamical approach (red solid lines in Fig. 7.9), striking improvements in skill can be achieved compared to the purely statistical approach. As can be seen from Fig. 7.9, the gain in skill decreases with the forecast week in both subregions as the valuable information contained in the initial conditions of the dynamical model becomes increasingly blurred. In comparison to the EPCopt model, the difference in the CRPSS decreases from $+17.2\%$ at week one to $+1.1\%$ at week three and vanishes thereafter in the Gulf of Mexico, whereas

it steadily decreases from $+9.2\%$ in week one but still exceeds the EPCopt model by $+1.1\%$ in week five. With the predictor pool only representing S2S-based predictions of the necessary environmental conditions for TC activity, but not including direct S2S predictions of ACE, the hybrid model approach is clearly worse than the S2SACE(-SPP) model at week one. But it outperforms all other approaches in the third week in the Gulf of Mexico, and from the second week on in the central MDR.

To assess the role of climatological information on TC occurrence, especially regarding its seasonal cycle, in comparison to the S2S-based predictors, a second version of the statistical-dynamical model is trained and validated without using the CSCopt predictor (red dotted lines in Fig. 7.9). Over the five forecast weeks considered, the CRPSS for this model version is decreased by only $0.3\,\%$ to $0.7\,\%$ in the Gulf of Mexico, whereas the reduction is in the range of $1.6\,\%$ to $2.3\,\%$ in the central MDR. From this different degree of drop in skill when removing the CSCopt predictor, it can be concluded that the hybrid model heavily builds on the enhanced predictability that is associated with the more pronounced seasonal cycle in the central MDR, which is also confirmed by the higher regression coefficients for the CSCopt predictor identified in that subregion (see Figs. 7.4b and 7.5b). Nevertheless, even without this valuable climatological information on TC occurrence, the information contained in the NWP-based predictors is sufficient to exceed the skill of all the other model approaches at week three in the Gulf of Mexico and week three and four in the central MDR, underpinning the great utility of the hybrid approach.

8. Conclusions

The typically extreme weather associated with TCs poses a particular threat to human life and infrastructure along affected coastlines. Early warnings are therefore necessary so that public decision-makers can take timely action and inform the population about evacuation plans. Operational forecasts typically focus on either short-to-medium range warnings or address seasonal predictions of integrated TC activity. Presumably due to lack of predictive skill in numerical forecast models, the subseasonal predictability gap has received increasing attention only in recent years. To contribute to closing the predictability gap with regard to TC activity, the aim of this dissertation was to identify and examine NWP-based predictors relevant for subseasonal TC activity forecasting, to develop and validate a statistical-dynamical forecasting model, and to systematically compare its predictive performance with a variety of distinct forecasting approaches. Following a review of commonly used modelling approaches for TC forecasting in Chapter 2, a hybrid model was developed for TC occurrence and ACE in Chapter 6 and Chapter 7, respectively, based on an extensive set of predictors that was developed and analysed in Chapter 5. This chapter summarizes the findings and answers the research questions posed in Chapter 3.

Ambient conditions supporting TC occurrence have been studied in detail for many decades, resulting in a commonly accepted list of environmental factors. However, it has not been analysed to date whether each of these factors remains relevant to subseasonal TC predictions when extracted from dynamical forecasts. On the other hand, only some studies set out to reveal potential links between different modes of oceanic and atmospheric variability, and subseasonal TC activity, inspired by statistical approaches commonly used for seasonal TC prediction. Chapter 5 aimed at combining the two approaches, namely to consider predictors representing ambient conditions as well as modulating effects from atmospheric modes of variability. Based on a literature review and physical considerations, an extensive predictor pool was generated comprising oceanic, tropical, and extratropical predictor groups. Predictor selection schemes were used for both target variables to extract predictor subsets for each forecast week and subregion, and the corresponding results allowed to assess relative predictor relevance. For TC oc-

currence, the successive inclusion of additional predictor groups provided further insight into how model skill builds up, which enabled the main part of the model skill to be attributed to individual predictor groups. Results of these investigations led to the following findings for research questions RQ1a-c:

RQ 1a ***What factors influencing TC occurrence are known in the literature, and are they likewise relevant when corresponding predictors are generated from subseasonal NWP forecasts?***

Previous approaches to develop subseasonal TC occurrence models used the seasonal cycle of TC occurrence probability as a base predictor (Leroy and Wheeler, 2008; Henderson and Maloney, 2013). Although North Atlantic SSTs closely follow this seasonal behavior, numerous studies point out the importance of local SSTs due to air-sea interactions (e.g., Palmen, 1948; Emanuel, 1986), as well as the circulation-altering role of remote SST patterns in the MDR or Niño regions (e.g., Shapiro, 1982; Goldenberg and Shapiro, 1996). The interplay of conducive ambient conditions is often assessed by indices such as the GPI (Emanuel and Nolan, 2004), whereas its components were used in this study to better link the target variables to physical drivers. While some studies have examined the modulating effects of tropical waves modes (e.g., Frank and Roundy, 2006), including the MJO (Maloney and Hartmann, 2000), on the TC environments, their relevance for subseasonal TC predictions needs further investigation. In recent years, increasing attention has been given in the literature to the identified link between the extratropical Rossby-wave breaking, represented by upper-level PV in this dissertation, and TC activity (e.g., Zhang et al., 2017a; Wang et al., 2020).

Using forecast week four to represent subseasonal timescales, an analysis of predictor-target correlation corroborated the basin-wide relevance of the climatological seasonal cycle. Local and MDR SSTs are similarly relevant with maximum positive correlations along the northern edge of the MDR region. In comparison, ENSO-related SSTs in the tropical Pacific are weaker correlated with a negative sign, with the Niño 1+2 region achieving the highest absolute correlations among all Niño regions. Significant correlations for the GPI components span from the West African coast to the Gulf of Mexico, along the classical track of TCs initiated by African Easterly waves, with a separate maximum in the western Gulf of Mexico. The sign of this correlation pattern is positive for 850-hPa absolute vorticity and 700-hPa rel-

ative humidity, and negative for 200-850-hPa vertical wind shear, which is thus consistent with the definition of the GPI. Tropical wave-filtered 200-hPa divergence squared yielded an activity measure for each wave type, which mostly correlates with a negative sign, except for a pronounced positive maximum at the northern edge of the central MDR. The upper-level PV predictor shows strong negative correlations in the northeastern edge of the MDR and in the western North Atlantic north of the MDR. Because each NWP-based predictor features significant correlation patterns with maximum values greater than 0.1 at week four, the assumption that they provide predictive subseasonal signals, useful as input for hybrid models, is verified.

RQ 1b ***What are the key predictors selected by the hybrid model at each forecast week when predicting TC occurrence and ACE, respectively?***

TC occurrence: Since the CSCopt base predictor was treated as fixed, it hence got the highest ranking in importance among all predictors by definition. Key predictors identified by the sequential predictor selection to be relevant at most lead times in the Gulf of Mexico are standard deviation of 850-hPa absolute vorticity and mean 700-hPa relative humidity. A striking transition occurs from week three to four, where the majority of tropical wave, GPI, and extratropical predictors is not selected at all. In the central MDR, mean 850-hPa absolute vorticity and mean 200-850-hPa vertical wind shear are found to be important predictors at all lead times. A noticeable transition occurs after week two, where local SST information and higher-frequency tropical wave modes (Kelvin and MRG/TD) become replaced by MDR SST and lower-frequency wave modes (MJO and ER), respectively.

ACE: In both subregions, no predictors were selected for the location model by the solver-integrated predictor selection, highlighting the challenge of predicting the distribution of weekly ACE and the need for a heteroscedastic model. For the scale model, mean MDR SST is found to be the overall most relevant predictor in both subregions, with an outstanding prominence in the Gulf of Mexico. In this subregion, apart from mean 200-500-hPa PV on the medium range and standard deviation of MDR SST at week three and four, no consistent patterns of higher absolute coefficients are identified over lead time. In the central MDR, mean Nino12 SST and mean local SST predictors contribute with higher absolute coefficients at subseasonal lead times. As in the TC occurrence model, the CSCopt predictor is found to play an important role at all lead times, due to the pronounced climatological seasonal

cycle of TC occurrence. While only a few single predictors were selected from the GPI and tropical wave groups, the mean low-frequency modes predictor appears to provide additional information that can be exploited for skill.

RQ 1c ***How much does each predictor group contribute to the predictive skill of the hybrid model?***

TC occurrence: When repeating the sequential predictor selection and training of the statistical–dynamical approach for CSCopt and GPI predictors only, the model is found to already perform better than the full predictor set provided to the purely statistical approach. A similar exceedance in skill occurs in the central MDR for the subseasonal lead times, when adding oceanic predictors first. Furthermore, the majority of the additional subseasonal skill stems from the GPI predictors in the Gulf of Mexico, and from the oceanic predictors in the central MDR. Tropical wave modes are found to have their strongest skill contribution at medium range.

Since the initial assumption that the environmental conditions are still useful when generating subseasonal predictors was verified, the second aim of this dissertation could be realized, namely the development of the statistical-dynamical hybrid model. The statistical component of the hybrid approach was implemented by a logistic regression model for TC occurrence and a two-part model model for the predictive distribution of ACE, respectively. Optimal subsets of predictors were determined for every forecast week and subregion by running predictor selection procedures, before the statistical models were used in forecast mode at every grid point separately. The fact that most predictors were actually selected during this procedure confirms their utility for the hybrid approach.

The third aim of this thesis was a systematic validation and comparison of the probabilistic forecasts from the different model types. While in the statistical–dynamical approach each predictor type was represented by the ensemble mean and standard deviation of S2S ECMWF reforecasts, an analogous set of predictors was derived from ERA5 data to train a purely statistical approach. The latter approach, already applied in previous studies (Leroy and Wheeler, 2008; Henderson and Maloney, 2013), served as reference to address the question whether extracting predictors from dynamical forecasts leads to better forecasting results. A comprehensive set of various benchmark models was created from climatologies and S2S ECMWF ensemble forecasts, which were further improved through optimization and calibration, respectively. The full set of model types has been validated for the Gulf of Mexico and the central MDR subregions separately, with the

following findings obtained regarding RQ2a-f. Note that results are valid for both target variables if not stated otherwise.

RQ 2a ***How well can each model discriminate between TC occurrence and non-occurrence, and are their forecasts calibrated?***

TC occurrence: Reliability diagrams indicate generally good calibrated models for forecast probabilities smaller than the mean relative frequency of the target variable, and increasing miscalibration at higher forecast probabilities. The calibration curve of the CSCopt model indicates an underforecasting bias, although climatological models should be calibrated by definition, as they are derived from past realizations of the observational distribution. Thus, this bias is most likely a result of non-stationarity, i.e., a trend between the full climatological reference and validation periods. Minor miscalibration is found for the statistical approaches, whereas the clear underforecasting bias present in the S2STC model can be corrected by IDR-based post-processing.

As diagnosed by means of BSS decomposition, the optimisation of the CSC model increases miscalibration but also leads to an even better discrimination between TC occurrence and non-occurrence. The best discrimination is given by the S2STC(cal) model at week one, but it quickly loses this ability over the medium range, discriminating worse than all the other model types on the subseasonal timescale. A comparison of the two statistical approaches further reveals that it is the NWP-based component of the hybrid model that improves both calibration and discrimination, particularly on the medium range.

ACE: Analogous to the results for TC occurrence prediction, analysis of uPIT histograms shows an underforecasting bias for the climatological and S2SACE models, which in case of the latter is corrected through IDR-based post-processing, while the two statistical approaches are already well calibrated.

RQ 2b ***How does the dynamical model perform over the five forecast weeks considered, and can statistical post-processing help improve the predictive performance?***

While the S2S ECMWF model predicts best at week one, it quickly drops in skill thereafter due to the chaotic nature of the atmosphere blurring the valuable information contained in the initial conditions. This considerable subseasonal loss in skill confirms the findings of previous studies (e.g., Lee

et al., 2018). For both target variables, the application of an IDR-based post-processing helps to raise skill at all lead times. However, the calibrated S2S model predictions do not exceed climatological and statistical models on subseasonal time scales.

RQ 2c ***In comparison, at which forecast week does the climatological model become more skillful, and is it worth optimizing its representation?***

Taking seasonality into account when building climatological models results in considerable improvements in predictive skill, in particular for the central MDR, which is subject to a more pronounced seasonal cycle in TC activity. On the other hand, larger skill improvements are found in the Gulf of Mexico when the seasonal cycle is optimized by incorporating climatological information from the right number of adjacent days. This suggests that using extension methods to populate the climatological data basis is particularly beneficial for subregions, where variations are less dominated by the seasonal cycle, and climatological models hence suffer from undersampling. The optimized climatological model already outperforms post-processed S2S forecasts from week three to four on. It therefore constitutes a good base predictor for the development of statistical models in the context of TC occurrence prediction.

RQ 2d ***Can the purely statistical modelling approach, using past data to generate predictors, exceed the skill of the climatological model at subseasonal lead times***

TC occurrence: The purely statistical approach from (Leroy and Wheeler, 2008; Henderson and Maloney, 2013), with logistic regression models trained on past data predictors, improves skill over the CSCopt model in both subregions up to week five. This demonstrates the ability of statistical models to add to the subseasonal skill present in climatological models if past data contain predictable, longer-lived signals.

ACE: In contrast, the purely statistical approach used in the context of the two-part model can only achieve slight improvements over the ECPopt model up to forecast week two to three. For subseasonal forecasting, this implies that data from before initialisation is not useful for the intensity-related predictive distribution of ACE, while it may provide a limited source of skill for TC occurrence prediction.

RQ 2e ***Does the statistical-dynamical approach, i.e., generating the same predictors from NWP forecasts, actually yield the putative subseasonal improvements?***

TC occurrence: With the statistical–dynamical approach, an even greater increase in model skill was found over the CSCopt model at all lead times considered, but especially on the medium range. Though this approach is still worse than the S2STCcal for week one, despite a significant increase in skill over the purely statistical approach, it outperforms all other models in the Gulf of Mexico from week three and in the central MDR from week two on. In both subregions, the subseasonal skill improvement from the purely statistical to the statistical–dynamical approach is up to about 4 times larger than the one from the CSCopt to the purely statistical approach. In view of the generally lower CSCopt skill in the Gulf of Mexico, such a relative improvement becomes even more remarkable, highlighting the value of this approach for subregions that are less subject to a seasonal cycle.

ACE: The benefit of using NWP-based predictors leads to an outstanding 17% increase in CRPSS compared to the EPCopt model at week one in the Gulf of Mexico, followed by a marked drop in improvements at week two, and completely vanishes from week four on. Although the week-one increase is smaller in the central MDR (up to 9%), clear improvements are gained at all lead times from the statistical-dynamical approach. The hybrid model is superior to all other models in the Gulf of Mexico at week three, and in the central MDR from week two on. An experiment withholding the CSCopt predictor resulted in a substantial decrease in skill at all lead times, demonstrating its major importance for the central MDR due to the inherent predictive information.

RQ 2f ***Which model provides the highest value for economic decision making at each forecast week?***

TC occurrence: The economic value of each calibrated model as function of a user's cost-loss ratio was determined based on the cost-loss decision model introduced by Richardson (2000, 2003). Regardless of the model or lead time considered, the maximum economic value that can be achieved is for users at a cost-loss ratio of 0.133 in the Gulf of Mexico and 0.118 in the central MDR. Generally speaking, a user at a cost-loss ratio of > 0.1 should consult S2STCcal predictions at week one and two in the Gulf of Mexico and at week one in the central MDR. For any other forecast week, the statistical-

dynamical hybrid model provides the most valuable information to reduce overall economic costs, as it mainly extends the value of climatological predictions at larger cost-loss ratios.

While the statistical-dynamical model approach has been used in other contexts before, its application for TC prediction has focused on either seasonal timescales (Klotzbach et al., 2019) or on prediction of subseasonal anomalies from climatology (Qian et al., 2020). The model development in the present dissertation is the first attempt to bridge the subseasonal predictability gap by a hybrid model that directly links dynamical predictors to weekly TC occurrence and the predictive distribution of ACE, respectively. Beyond the mere development of another - though well-founded - model approach, the strength of the study lies in the subsequent systematic and comprehensive validation of a hierarchy of distinct model approaches. Starting from original approaches, such as a plain mean climatology or raw S2S model output, more sophisticated models were developed and added to the set of benchmark models, e.g. by representing the seasonal cycle, optimizing the population strategy of the climate data basis, or statistical post-processing. The demonstrated value of the efforts made to improve the benchmark models is two-fold. It puts the newly developed hybrid model into strong and thus honest competition, and on the other hand provides a more solid basis for validation, which strengthens findings and conclusions. Methodically, established forecast verification tools were complemented by newly developed techniques (see Sections 4.4 and 4.5), that either solved long-standing problems (e.g. CORP approach providing optimal bins) or generalized concepts for new fields of application (e.g. UROC curve).

The systematic comparison of original and derived model types presented has demonstrated the great potential of statistical–dynamical modeling for a specific application of extreme events on the subseasonal forecast horizon. Exploiting S2S forecasts to develop a hybrid model proved to be the best strategy - at present - for probabilistic forecasting of subregional North Atlantic TC occurrence and the predictive distribution of ACE beyond week one to two, respectively, and may be a promising strategy for other (sub)basins and forecasting applications as well. Despite the identified improvements in forecast skill, predicting TC activity remains highly challenging, especially on subseasonal lead times.

9. Outlook

To provide warnings of TC-related hazards beyond the usual operational forecast lead times, a statistical-dynamical hybrid model was created to forecast TC occurrence and ACE, and validated against a hierarchy of models. Despite the effort that went into model development and forecast evaluation, some approaches and ideas, that were beyond the scope of this dissertation, as well as limitations are presented in the following and are left for further research.

With the spatio-temporal aggregation used to define the two target variables, the formulation of the forecasting problem is beyond the level of individual TCs, but more restricted compared to the basin-wide evaluation typically used for seasonal outlooks. Instead of a large set of target variables, the focus was laid on a comprehensive and systematic model validation that enabled the assessment of a hierarchy of different forecasting approaches. It should be noted, however, that the findings in terms of relative model performance cannot bet readily transferred to differing target variable definitions. A clear consensus in the TC research and forecasting community, about which variables to look at and how they should be defined, does not exist yet (Camargo et al., 2019). Therefore, numerous distinct approaches have been considered for this model comparison, to bridge the time until such a consensus is hopefully reached, and new studies can more directly build on previous results.

As necessary as conventions are, they yet can pose man-made problems that do not necessarily occur in nature. Following a commonly accepted convention, cyclones were only considered if they exceeded tropical storm strength ($\geq 34\,\mathrm{kt}$). Although this threshold may be well justified empirically, it introduces an artificial separation between occurrence and non-occurrence, whereas nature produces cyclones on a continuous scale and does not care about any separation. The introduction of such a threshold thus renders the original continuous problem into a binary one, which then necessitates probabilistic forecasts to bring back continuity between the two classes. A better approach to formulate the problem would be to avoid any threshold and use a continuous target variable instead. However, this would require an observational dataset that does not yet exist, including the parts of the tracks where the cyclones are weaker than tropical storm strength.

The models were developed for use throughout the full hurricane season. Even though a predictor was provided to represent the seasonal cycle, it may be worth to tailor the statistical models to predict for specific sub-periods (e.g., early, peak, and late season). That way, characteristic intraseasonal variations inherent in predictor variables could be better exploited, in particular in subregions with a less pronounced seasonal cycle, such as the Gulf of Mexico. A larger dataset than the one available in this thesis would certainly be necessary for this approach, which is why it was not applied here.

When predicting on subseasonal timescales, the question quickly arises at what lead time statistical models trained on slow-varying predictors are more skillful than dynamical models. However, the systematic comparison of approaches highlighted the great potential of climatological models that, in the author's opinion, are often underrated. The fact that a plain climatological model performed better than a statistically post-processed state-of-the-art NWP model is remarkable, and this approach should hence be considered for other subseasonal prediction tasks as well. Before making the effort to design any statistical or statistical-dynamical model for subseasonal predictions, it may be worth creating and further optimizing climatological models, since they need to be computed only once and with orders of magnitude lower computational costs.

The statistical-dynamical hybrid approach developed in this dissertation for TC occurrence and ACE distribution was superior in terms of subseasonal predictive skill compared to all other models. Even though the utility of this approach has been clearly demonstrated for TC forecasting, improvements and further developments can be made to various parts and at different levels. The highest level would be to separate between changes in the dynamical and the statistical component, which - either alone or in combination - can be refined or even replaced by more sophisticated approaches to further leverage predictive power. Regarding the dynamical model part, more predictor types and variables could be added to the pool, and/or their ensemble distribution could be represented in ways other than by its mean and standard deviation. Moreover, predictors could be constructed from other S2S models, and/or with potentially higher spatio-temporal resolution if available in the future. The mean bias correction applied to the underlying variables of the NWP-based predictors could be replaced by quantile mapping, because forecast errors are probably different from the mean bias if extreme values are predicted.

As for the statistical component, the logistic regression and two-part models could be replaced by other parametric, but also nonparametric, modelling approaches. Because TC occurrence is associated with several nonlinear processes (e.g., convection, tropical wave interaction, or extratropical Rossby-wave breaking), a considerable part of the work leading to this dissertation was dedicated to also test the predictive performance of deep learning approaches (LeCun et al., 2015), that were expected to be able to skilfully model such nonlinear dependencies. Having tested different neural network (NN)

architectures, choices for hyperparameters, and sub-regional pooling of data to increase training datasets, NN approaches turned out to not perform any better than the logistic regression models for TC occurrence. This is despite the fact that quite some effort has been made in terms of model regularization, i.e., to prevent models from over- and underfitting. Due to the data-driven learning and the concomitant limitations for exploration of decision making, definite conclusions as to why the NNs were unable to further leverage skill could not be drawn. However, it seems that the low signal-to-noise ratio in the subseasonal ensemble forecasts and the generally low relative frequency of TC occurrence prevent the NNs from unfolding their capability of modelling nonlinearities. To test whether deep learning can identify synoptic-to-planetary scale patterns as additional source of subseasonal information, convolutional NNs (Dhillon and Verma, 2020) were trained on the underlying dynamical forecast fields directly, thereby incorporating predictor generation into the modelling process. This approach, however, did not gain any improvements over the logistic regression approach either. At this stage, it remains unclear whether deep learning models have the potential to improve subseasonal forecasts for TC activity. A thorough investigation of deep learning models was beyond the scope of this dissertation, and thus future research would be needed in this direction.

With its focus on the North Atlantic, this thesis validated the hierarchy of models for a basin, in which actual and potential NWP model skill for predicting TC occurrence are close together (Lee et al., 2018). While this fact had some advantages when drawing conclusions about relative improvements in skill, the question whether the hybrid approach can lead to even higher relative improvements in other basins, where NWP models have still great potential for improvement (e.g., the southern Indian Ocean), remains open. Further research is needed to improve the utility of the statistical-dynamical hybrid approach and test it for other oceans. Confirming the finding of Vitart et al. (2010), first promising results have shown that averaging methods can yield further slight improvements, as combining predictions from different model types typically reduces variance.

A. Appendix

A.1 Model comparison for TC occurrence

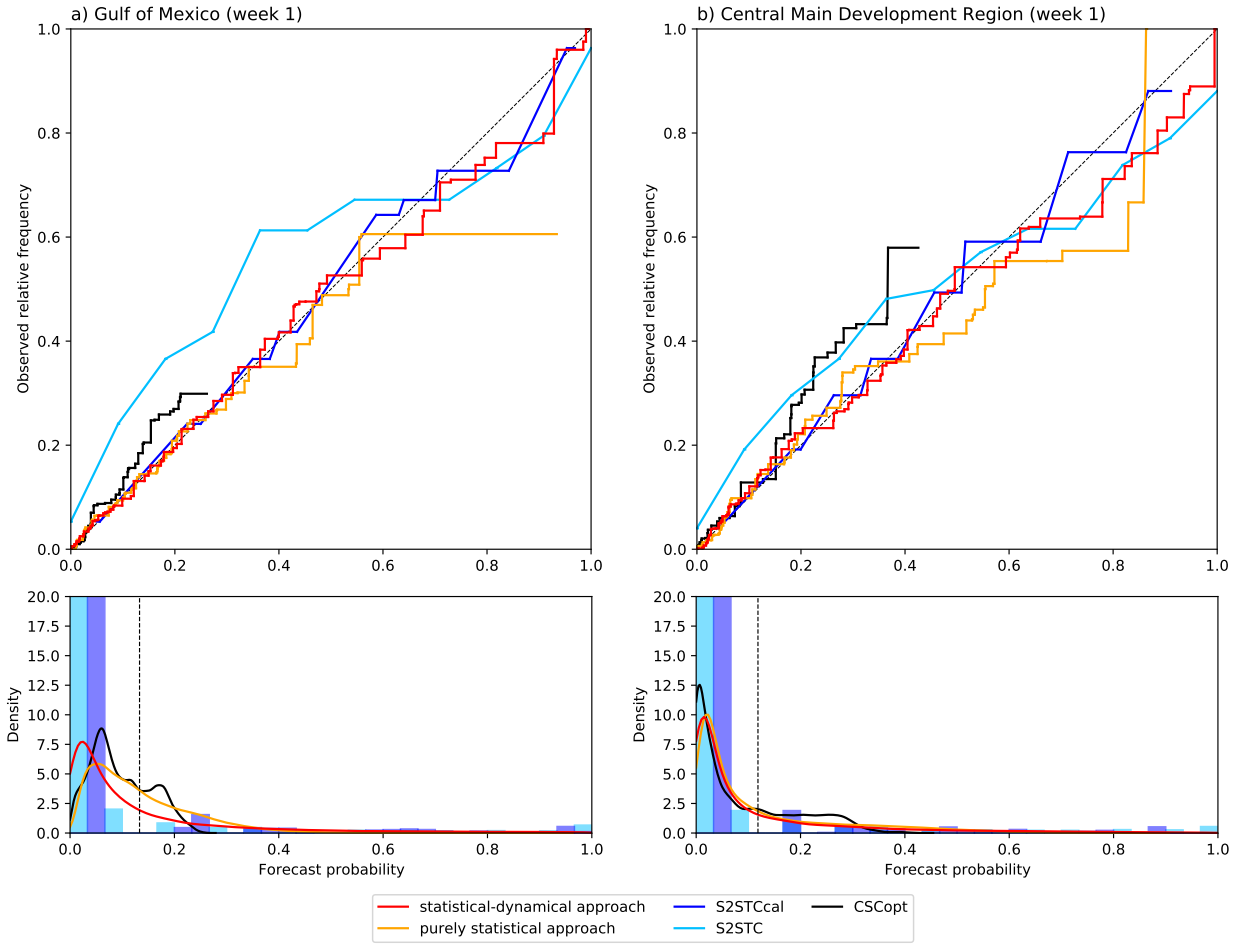

Figure A.1.1: Same as Fig. 6.6, but calculated for week 1.

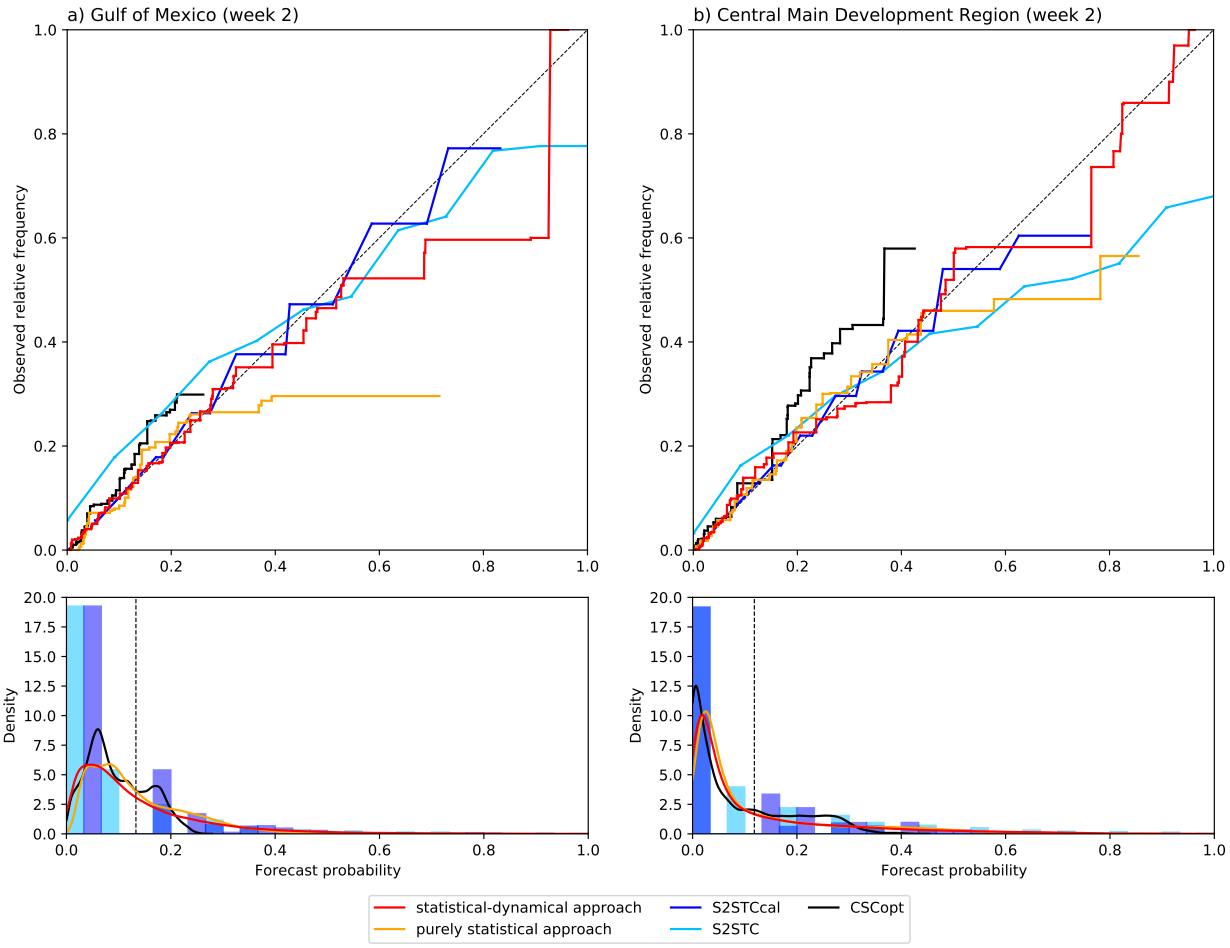

Figure A.1.2: Same as Fig. 6.6, but calculated for week 2.

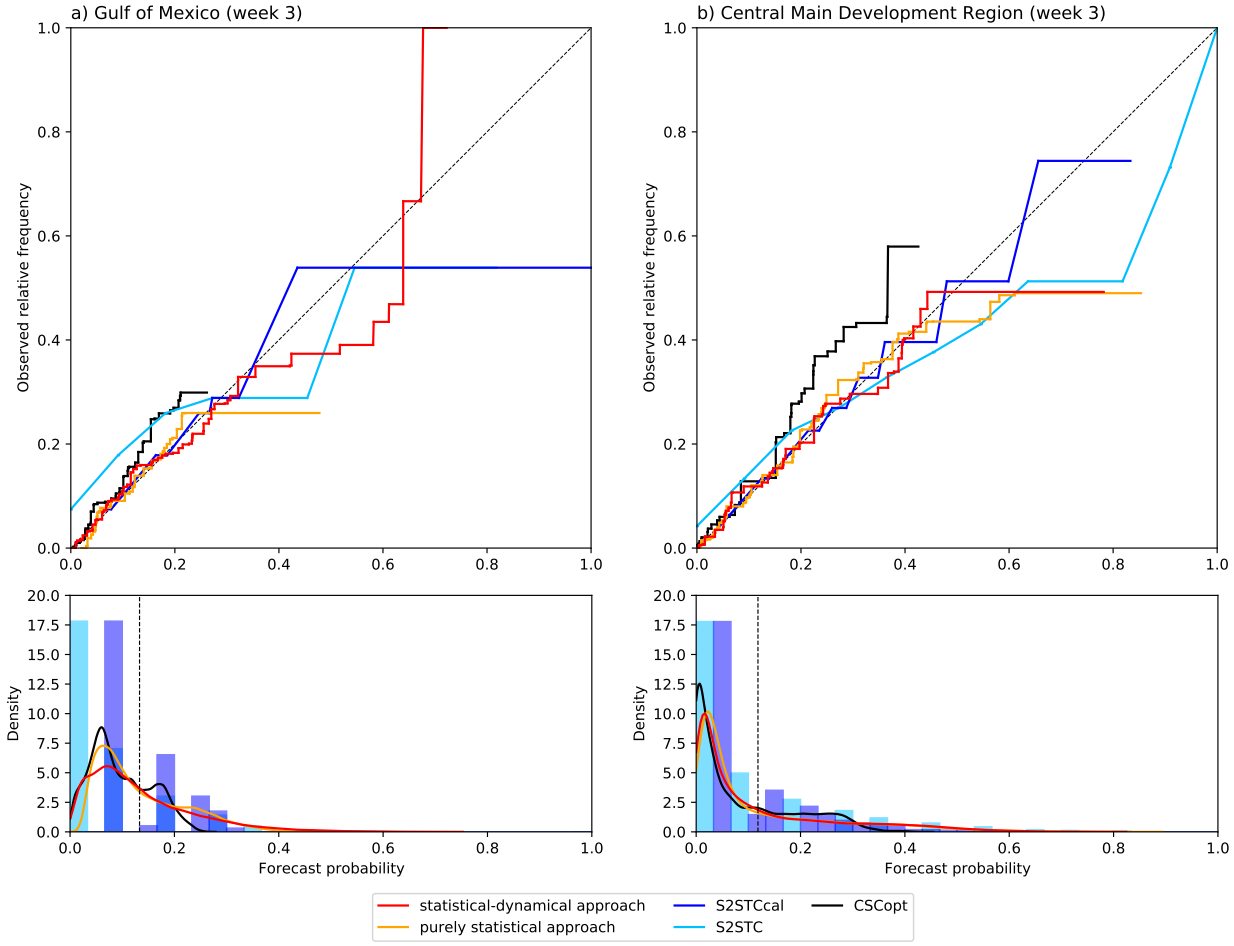

Figure A.1.3: Same as Fig. 6.6, but calculated for week 3.

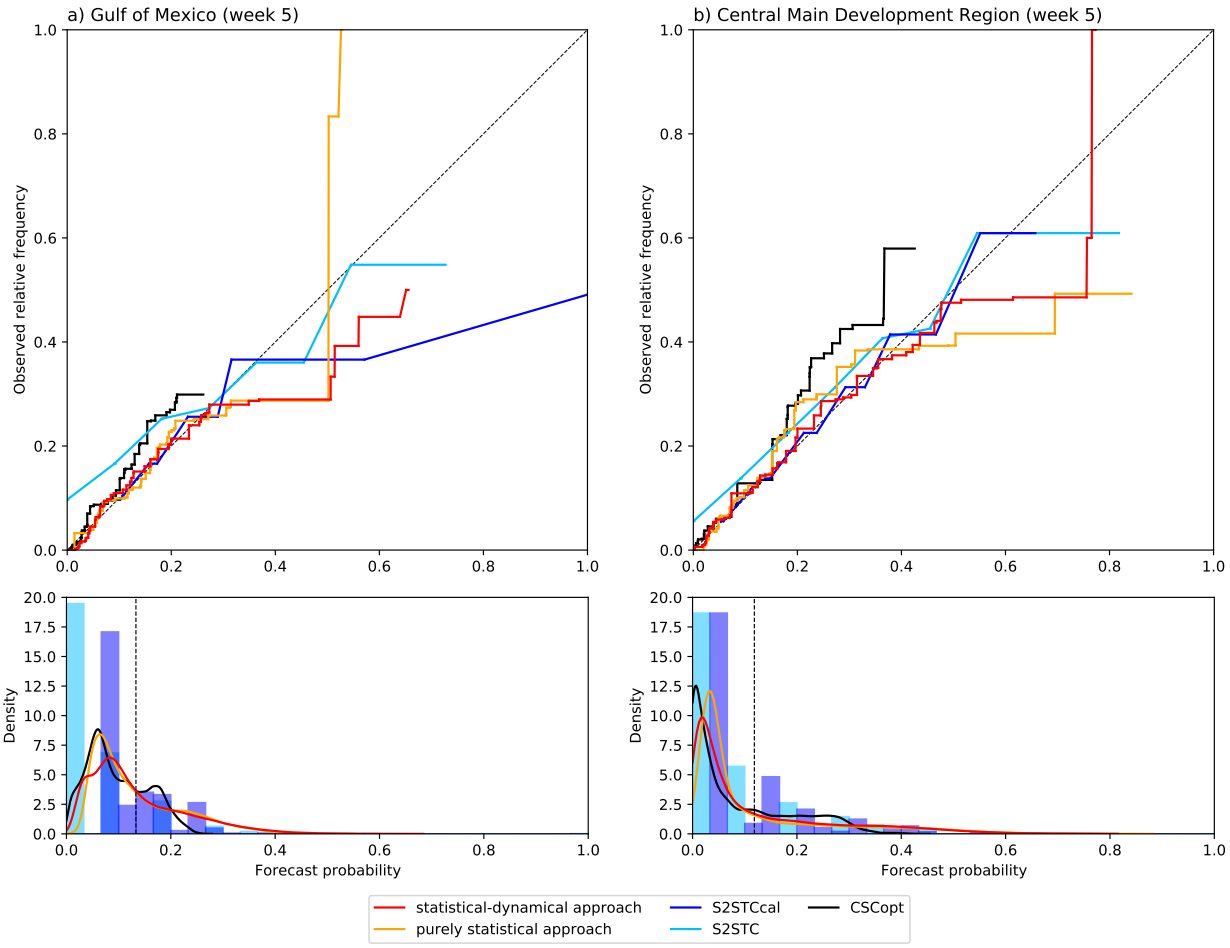

Figure A.1.4: Same as Fig. 6.6, but calculated for week 5.

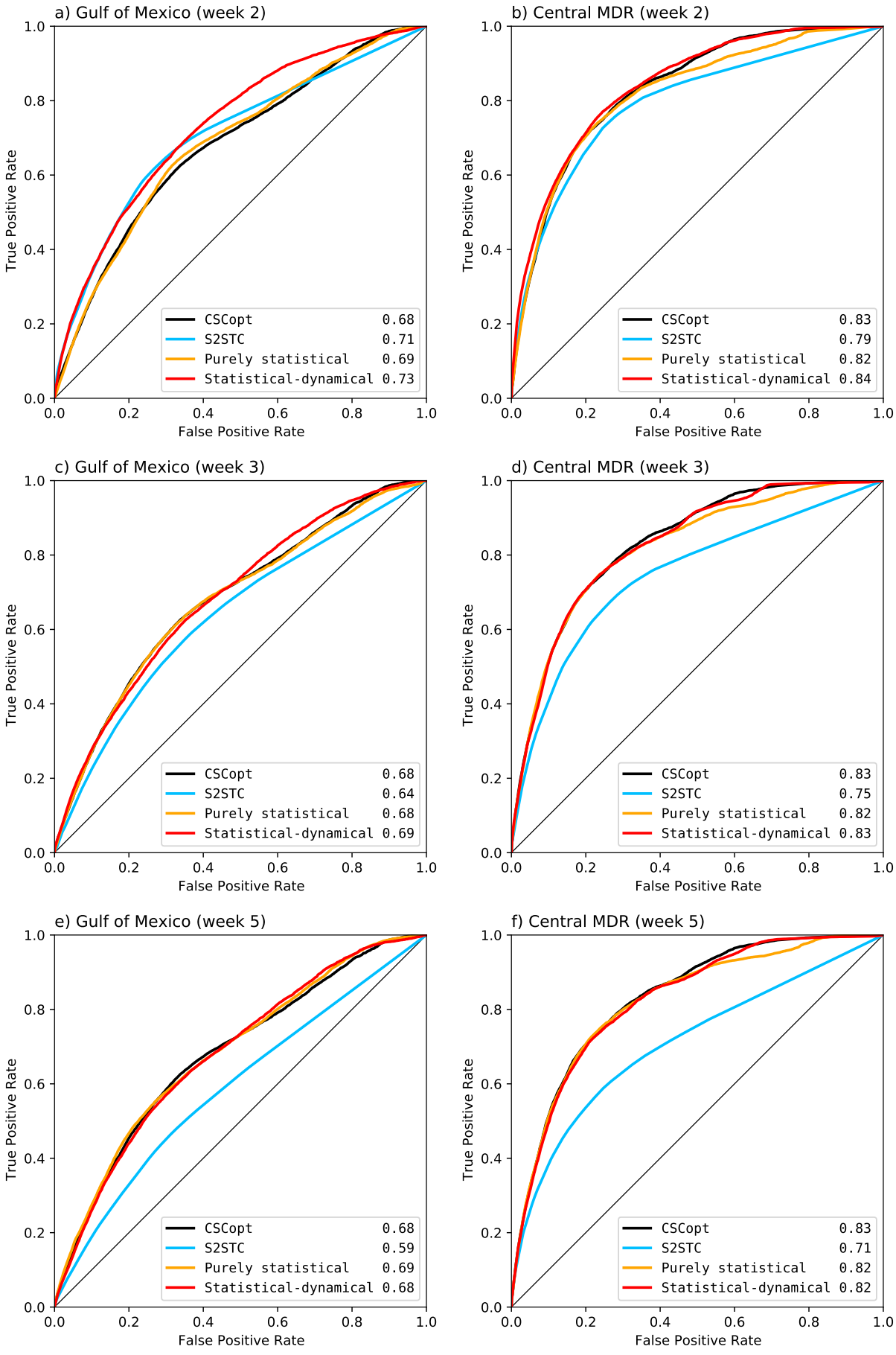

Figure A.1.5: Same as Fig. 6.7, but calculated for (a+b) week 2, (c+d) week 3, and (e+f) week 5.

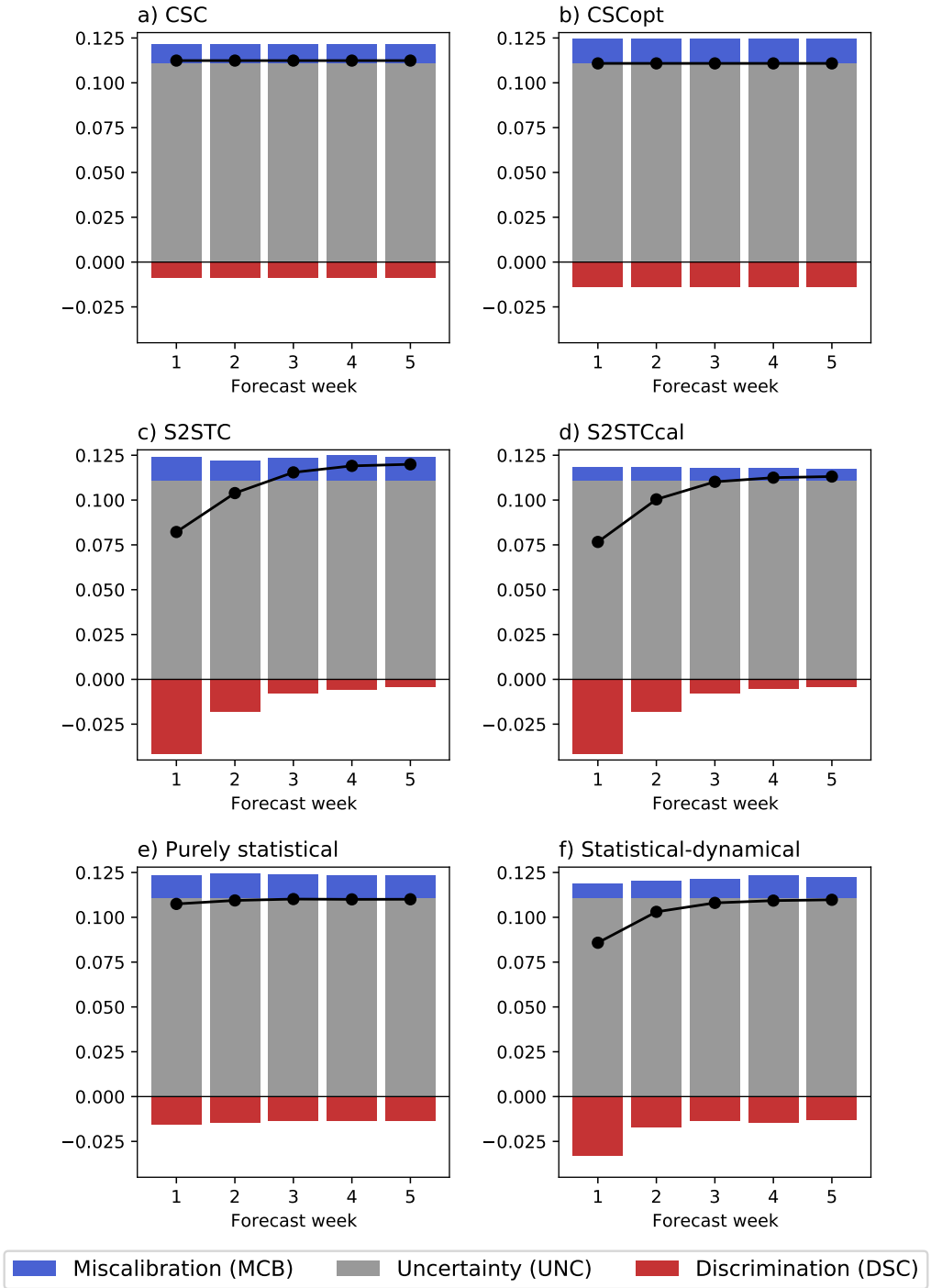

Figure A.1.6: Same as Fig. 6.9, but for the Gulf of Mexico subregion.

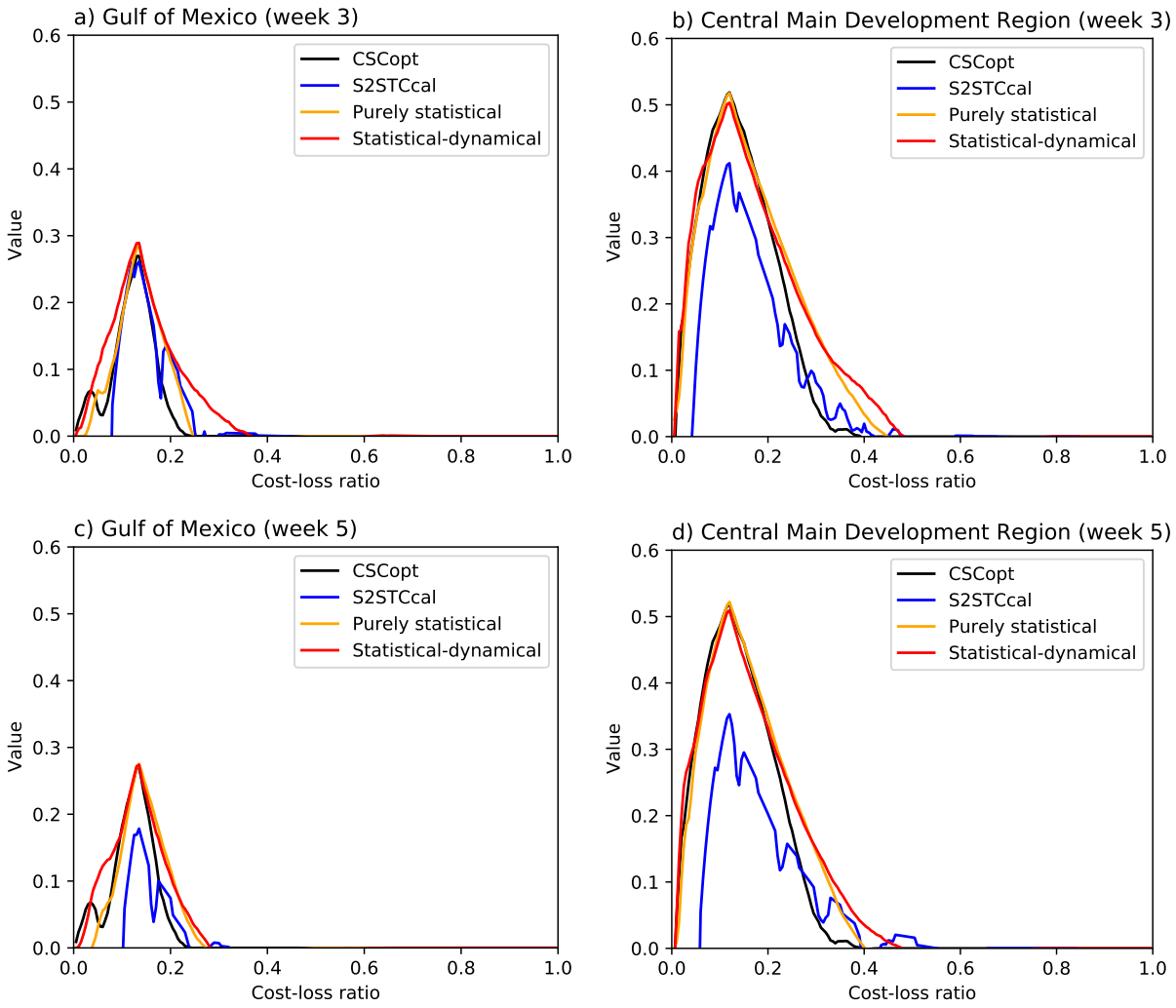

Figure A.1.7: Same as Fig. 6.11, but calculated for (a+b) week 3, and (c+d) week 5.

A.2 Model comparison for accumulated cyclone energy

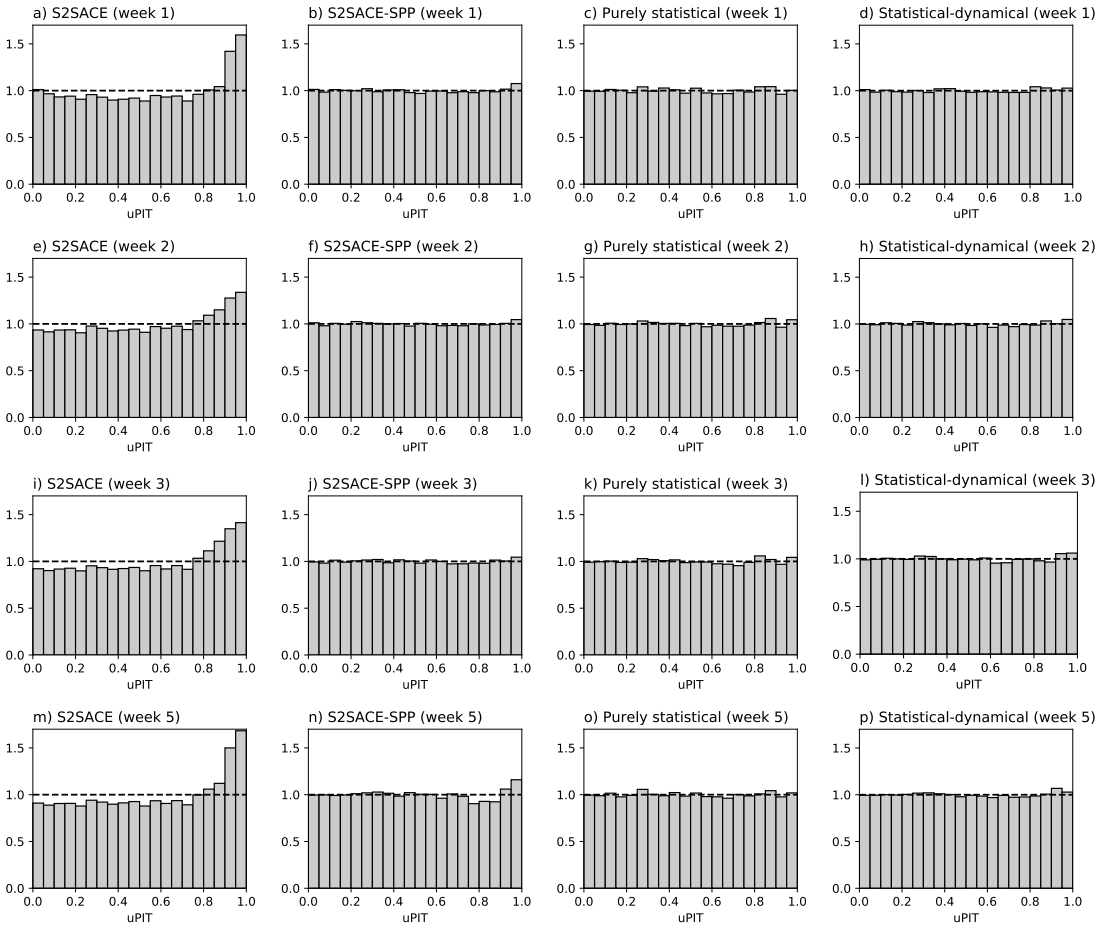

Figure A.2.1: Same as Fig. 7.6, but for (a,e,i,m) S2SACE, (b,f,j,n) S2SACE-SPP, (c,g,k,o) purely statistical, (d,h,l,p) statistical-dynamical forecasts at week (a-d) one, (e-h) two, (i-l) three, and (m-p) five.

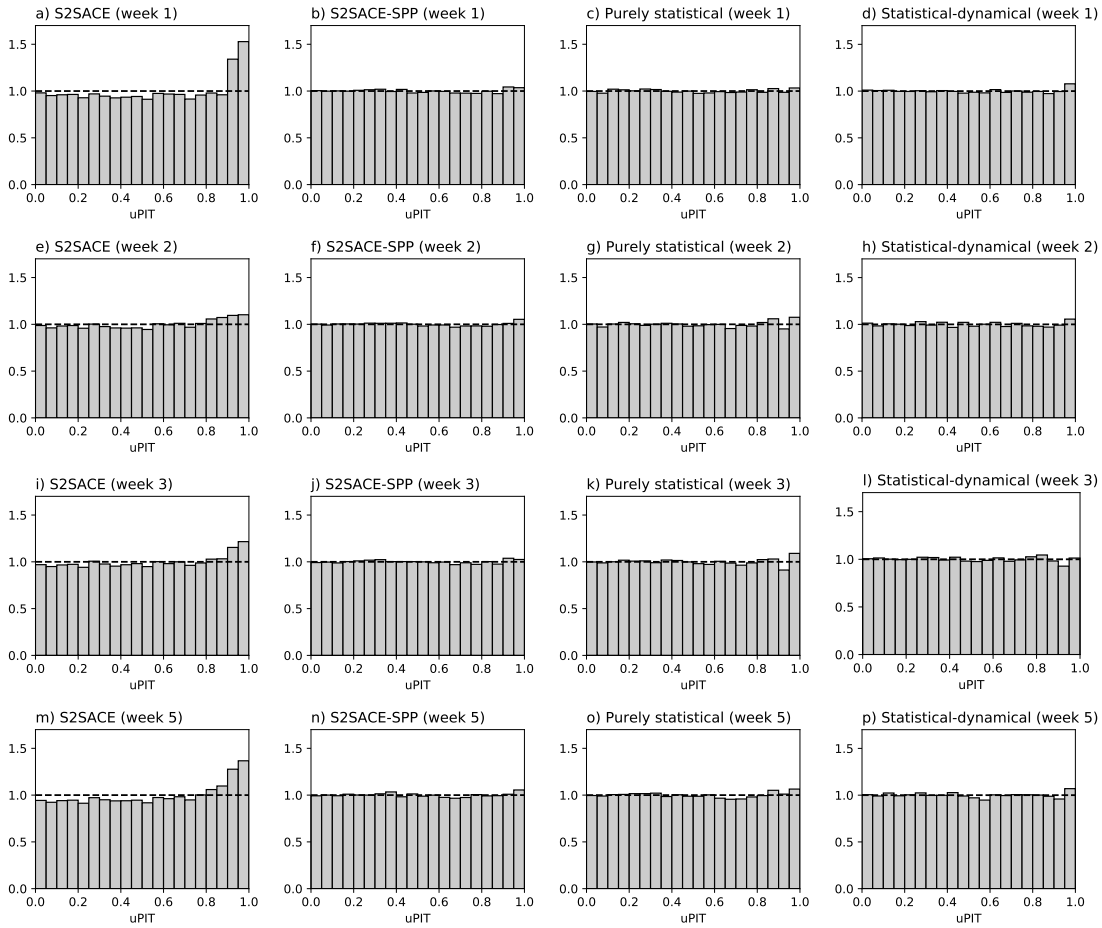

Figure A.2.2: As in Fig. 7.7, but for (a,e,i,m) S2SACE, (b,f,j,n) S2SACE-SPP, (c,g,k,o) purely statistical, (d,h,l,p) statistical-dynamical forecasts at week (a-d) one, (e-h) two, (i-l) three, and (m-p) five.

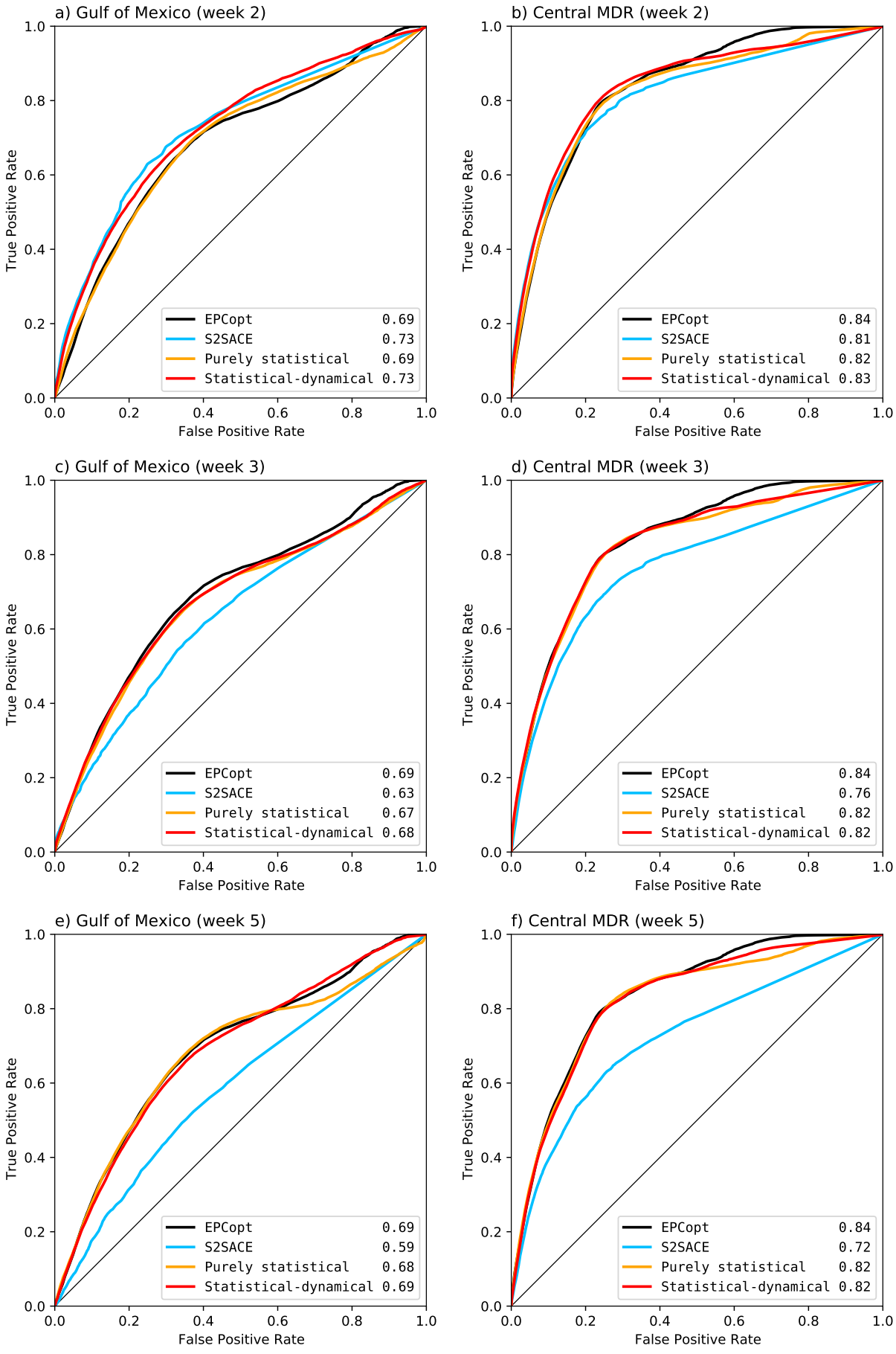

Figure A.2.3: Same as Fig. 7.8, but calculated for (a+b) week 2, (c+d) week 3, and (e+f) week 5.

Bibliography

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

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, **9 (7)**, 1518–1530.

Ångström, A., 1922: *On the effectivity of weather warnings*, Vol. 1. Nordisk Statistisk Tidskrift, 394-408 pp.

AOML, 2021: Hurricanes Frequently Asked Questions. Accessed 27 September 2021. Available online at https://www.aoml.noaa.gov/hrd-faq/#record-setters.

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

Belanger, J. I., P. J. Webster, J. A. Curry, and M. T. Jelinek, 2012: Extended prediction of north indian ocean tropical cyclones. *Weather and Forecasting*, **27 (3)**, 757–769, doi:10.1175/WAF-D-11-00083.1.

Bell, G. D., and Coauthors, 2000: Climate Assessment for 1999 Table of Contents. *Bulletin of the American Meteorological Society*, **81 (6)**, 1–50, doi:10.1175/1520-0477(2000)81[s1:CAF]2.0.CO;2.

Bister, M., and K. A. Emanuel, 1998: Dissipative heating and hurricane intensity. *Meteorology and Atmospheric Physics*, **65 (3-4)**, 233–240, doi:10.1007/BF01030791.

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. *Monthly Weather Review*, **78**, 1–3, doi:10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Bühlmann, P., and T. Hothorn, 2007: Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, **22 (4)**, 477–505, doi:10.1214/07-STS242.

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ecmwf ensemble prediction system. *Monthly Weather Review*, **125 (1)**, 99–119, doi:10.1175/1520-0493(1997)125⟨0099:PFSOEP⟩2.0.CO;2.

Camargo, S. J., A. G. Barnston, P. J. Klotzbach, and C. W. Landsea, 2007: Seasonal tropical cyclone forecasts. *WMO Bulletin*, **56 (4)**, 297–309.

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

Camargo, S. J., and Coauthors, 2019: Tropical cyclone prediction on subseasonal timescales. *Tropical Cyclone Research and Review*, **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. *Quarterly Journal of the Royal Meteorological Society*, **144**, 1337–1351, doi:10.1002/qj.3260.

Caron, L.-P., L. Hermanson, and F. J. Doblas-Reyes, 2015: Multiannual forecasts of atlantic u.s. tropical cyclone wind damage potential. *Geophysical Research Letters*, **42 (7)**, 2417–2425, doi:https://doi.org/10.1002/2015GL063303.

CIRA RAMMB, 2022: Regional and Mesoscale Meteorology Branch (RAMMB) Statistical Tropical Cyclone Intensity Forecast Technique Development. Accessed 1 November 2022. Available online at https://rammb2.cira.colostate.edu/research/tropical-cyclones/ships/.

Cragg, J. G., 1971: Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, **39 (5)**, 829–844, doi:10.2307/1909582.

Davis, C. A., and L. F. Bosart, 2003: Baroclinically induced tropical cyclogenesis. *Monthly Weather Review*, **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. *Bulletin of the American Meteorological Society*, **85**, 1657–1662, doi:10.1175/BAMS-85-11-1657.

DeMaria, M., and J. Kaplan, 1994: A statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic basin. *Weather and Forecasting*, **9 (2)**, 209–220, doi:10.1175/1520-0434(1994)009⟨0209:ASHIPS⟩2.0.CO;2.

DeMaria, M., M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan, 2005: Further improvements to the statistical hurricane intensity prediction scheme (SHIPS). *Weather and Forecasting*, **20 (4)**, 531–543, doi:10.1175/WAF862.1.

Dhillon, A., and G. K. Verma, 2020: Convolutional neural network: a review of models, methodologies and applications to object detection. *Progress in Artificial Intelligence*, **9 (2)**, 85–112, doi:10.1007/s13748-019-00203-0.

Dimitriadis, T., T. Gneiting, and A. I. Jordan, 2021: Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, **118 (8)**, e2016191 118, doi:10.1073/pnas.2016191118.

Doms, G., and M. Baldauf, 2021: A description of the nonhydrostatic regional COSMO-Model. Part I: Dynamics and numerics. Consortium for Small-Scale Modelling, 168 pp., doi:10.5676/dwd_pub/nwv/cosmo-doc_6.00_I.

Doms, G., and Coauthors, 2021: A description of the nonhydrostatic regional COSMO-Model. Part II: Physical parameterizations. Consortium for Small-Scale Modelling, 177 pp., doi:10.5676/dwd_pub/nwv/cosmo-doc_6.00_II.

Doyle, J. D., and Coauthors, 2014: Tropical cyclone prediction using coamps-tc. *Oceanography*, **27 (3)**, 104–115.

Drews, C., 2007: Separating the ACE hurricane index into number, intensity, and duration. Accessed 12 August 2021. Available online at https://acomstaff.acom.ucar.edu/drews/hurricane/SeparatingTheACE.html.

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

Dunnavan, G. M., and J. W. Diercks, 1980: An analysis of super typhoon Tip (October 1979). *Monthly Weather Review*, **108 (11)**, 1915–1923, doi:10.1175/1520-0493(1980)108⟨1915:AAOSTT⟩2.0.CO;2.

Egan, J. P., G. Z. Greenberg, and A. I. Schulman, 1961: Operating characteristics, signal detectability, and the method of free response. *The Journal of the Acoustical Society of America*, **33 (8)**, 993–1007.

Emanuel, K. A., 1986: An air-sea interaction theory for tropical cyclones. Part I: Steady-state maintenance. *Journal of the Atmospheric Sciences*, **43**, 585–605, doi:10.1175/1520-0469(1986)043⟨0585:AASITF⟩2.0.CO;2.

Emanuel, K. A., 1999: The power of a hurricane: An example of reckless driving on the information superhighway. *Weather*, **54 (4)**, 107–108, doi:10.1002/j.1477-8696.1999.tb06435.x.

Emanuel, K. A., 2005: Increasing destructiveness of tropical cyclones over the past 30 years. *Nature*, **436 (7051)**, 686–688, doi:10.1038/nature03906.

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.].

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, **8 (6)**, 985–987.

Evans, C., and Coauthors, 2017: The extratropical transition of tropical cyclones. Part I: Cyclone evolution and direct impacts. *Monthly Weather Review*, **145 (11)**, 4317–4344, doi:10.1175/MWR-D-17-0027.1.

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

Fawcett, T., 2006: An introduction to roc analysis. *Pattern Recognition Letters*, **27 (8)**, 861–874.

Fox, K. R., and F. Judt, 2018: A numerical study on the extreme intensification of hurricane patricia (2015). *Weather and Forecasting*, **33 (4)**, 989–999, doi:10.1175/WAF-D-17-0101.1.

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

Fujiwhara, S., 1923: On the growth and decay of vortical systems. *Quarterly Journal of the Royal Meteorological Society*, **49 (206)**, 75–104, doi:https://doi.org/10.1002/qj.49704920602.

Fujiwhara, S., 1931: Short note on the behavior of two vortices. *Proceedings of the Physico-Mathematical Society of Japan. 3rd Series*, **13 (3)**, 106–110, doi:10.11429/ppmsj1919.13.3_106.

Gall, R., J. Franklin, F. Marks, E. N. Rappaport, and F. Toepfer, 2013: The hurricane forecast improvement project. *Bulletin of the American Meteorological Society*, **94 (3)**, 329–343, doi:10.1175/BAMS-D-12-00071.1.

Gentry, M. S., and G. M. Lackmann, 2010: Sensitivity of simulated tropical cyclone structure and intensity to horizontal resolution. *Monthly Weather Review*, **138 (3)**, 688–704.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69 (2)**, 243–268.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102 (477)**, 359–378, doi: 10.1198/016214506000001437.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133 (5)**, 1098–1118.

Gneiting, T., and P. Vogel, 2018: Receiver operating characteristic (ROC) curves. *arXiv preprint arXiv:1809.04808*.

Gneiting, T., and P. Vogel, 2019: Receiver operating characteristic (ROC) curves: What are they and what are they good for? *39th International Symposium on Forecasting*, Thessaloniki, International Institute of Forecasters, [Available online at https://isf.forecasters.org/wp-content/uploads/gravity_forms/2-dd30f7ae09136fa695c552259bdb3f99/2019/07/Tilmann_Gneiting_ISF2019.pdf.].

Gneiting, T., and E.-M. Walz, 2019: Receiver operating characteristic (ROC) movies, universal ROC (UROC) curves, and coefficient of predictive ability (CPA). *arXiv preprint arXiv:1912.01956*.

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. *Journal of 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. *Monthly Weather Review*, **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. *Monthly Weather Review*, **112**, 1649–1668, doi: 10.1175/1520-0493(1984)112⟨1649:ASHFPI⟩2.0.CO;2.

Gray, W. M., C. W. Landsea, P. W. Mielke, K. J. Berry, W. M. Gray, C. W. Landsea, P. W. M. Jr., and K. J. Berry, 1992: Predicting Atlantic Seasonal Hurricane Activity 6–11 Months in Advance. *Weather and Forecasting*, **7 (3)**, 440–455, doi: 10.1175/1520-0434(1992)007⟨0440:PASHAM⟩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. *Atmospheric Science Letters*, **20 (4)**, e886, doi:10.1002/asl.886.

Guo, C., G. Pleiss, Y. Sun, and K. Q. Weinberger, 2017: On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, PMLR, Proceedings of Machine Learning Research, Vol. 70, 1321–1330.

Hall, T. M., and S. Jewson, 2007: Statistical modelling of North Atlantic tropical cyclone tracks. *Tellus A: Dynamic Meteorology and Oceanography*, **59 (4)**, 486–498, doi:10.1111/j.1600-0870.2007.00240.x.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, **129 (3)**, 550–560.

Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Monthly Weather Review*, **125 (6)**, 1312–1327.

Harper, B., J. Kepert, and J. Ginger, 2010: *Guidelines for converting between various wind averaging periods in tropical cyclone conditions*. WMO/TD 1555, World Meteorological Organization, 54 pp., accessed 7 November 2022. Available online at https://library.wmo.int/doc_num.php?explnum_id=290.

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.

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

Henzi, A., J. F. Ziegel, and T. Gneiting, 2021: Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **83 (5)**, 963–993, doi:10.1111/rssb.12450.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15 (5)**, 559–570.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **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. *Monthly Weather Review*, **146**, 2337–2360, doi:10.1175/MWR-D-17-0261.1.

Jolliffe, I. T., and D. B. Stephenson, 2012: *Forecast verification: A practitioner's guide in atmospheric science*. John Wiley & Sons.

Judt, F., and Coauthors, 2021: Tropical Cyclones in Global Storm-Resolving Models. *Journal of the Meteorological Society of Japan. Ser. II*, **99**, 579–602, doi:10.2151/jmsj.2021-029.

Kaplan, J., M. DeMaria, and J. A. Knaff, 2010: A revised tropical cyclone rapid intensification index for the atlantic and eastern north pacific basins. *Weather and Forecasting*, **25 (1)**, 220–241, doi:10.1175/2009WAF2222280.1.

Keller, J. H., and Coauthors, 2019: The extratropical transition of tropical cyclones. Part II: Interaction with the midlatitude flow, downstream impacts, and implications for predictability. *Monthly Weather Review*, **147 (4)**, 1077–1106, doi:10.1175/MWR-D-17-0329.1.

Kim, H., F. Vitart, and D. E. Waliser, 2018: Prediction of the madden–julian oscillation: A review. *Journal of Climate*, **31 (23)**, 9425–9443, doi:10.1175/JCLI-D-18-0210.1.

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

Klotzbach, P., and Coauthors, 2019: Seasonal tropical cyclone forecasting. *Tropical Cyclone Research and Review*, **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. *Journal of Climate*, **27**, 2317–2330, doi:10.1175/JCLI-D-13-00483.1.

Klotzbach, P. J., M. A. Saunders, G. D. Bell, and E. S. Blake, 2017: *North Atlantic Seasonal Hurricane Prediction*, chap. 19, 315–328. American Geophysical Union (AGU), doi:https://doi.org/10.1002/9781119068020.ch19.

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. *Bulletin of the American Meteorological Society*, **91**, 363–376, doi:10.1175/2009BAMS2755.1.

Kossin, J. P., and M. Sitkowski, 2012: Predicting hurricane intensity and structure changes associated with eyewall replacement cycles. *Weather and Forecasting*, **27 (2)**, 484–488, doi:10.1175/WAF-D-11-00106.1.

Krishnamoorthy, K., 2006: *Handbook of statistical distributions with applications*. Chapman and Hall/CRC.

Landsea, C. W., and J. P. Cangialosi, 2018: Have we reached the limits of predictability for tropical cyclone track forecasting? *Bulletin of the American Meteorological Society*, **99 (11)**, 2237–2243, doi:10.1175/BAMS-D-17-0136.1.

LeCun, Y., Y. Bengio, and G. Hinton, 2015: Deep learning. *Nature*, **521 (7553)**, 436–444, doi:10.1038/nature14539.

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. *Weather and 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. *Weather and 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. *Monthly Weather Review*, **136**, 3637–3654, doi:10.1175/2008MWR2426.1.

Leutbecher, M., and T. Palmer, 2008: Ensemble forecasting. *Journal of Computational Physics*, **227 (7)**, 3515–3539, doi:10.1016/j.jcp.2007.02.014.

Madden, R. A., and P. R. Julian, 1971: Detection of a 40–50 day oscillation in the zonal wind in the tropical pacific. *Journal of Atmospheric Sciences*, **28 (5)**, 702–708, doi:10.1175/1520-0469(1971)028⟨0702:DOADOI⟩2.0.CO;2.

Madden, R. A., and P. R. Julian, 1972: Description of global-scale circulation cells in the tropics with a 40–50 day period. *Journal of Atmospheric Sciences*, **29 (6)**, 1109–1123, doi:10.1175/1520-0469(1972)029⟨1109:DOGSCC⟩2.0.CO;2.

Maier-Gerber, M., A. H. Fink, M. Riemer, E. Schoemer, C. Fischer, and B. Schulz, 2021: Statistical–dynamical forecasting of subseasonal north atlantic tropical cyclone occurrence. *Weather and Forecasting*, **36 (6)**, 2127–2142, doi:10.1175/WAF-D-21-0020.1.

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. *Monthly Weather Review*, **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.

Mason, I., 1982: A model for assessment of weather forecasts. *Australian Meteorological Magazine*, **30 (4)**, 291–303.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Management Science*, **22 (10)**, 1087–1096, doi:10.1287/mnsc.22.10.1087.

McTaggart-Cowan, R., G. D. Deane, L. F. Bosart, C. A. Davis, and T. J. Galarneau Jr, 2008: Climatology of tropical cyclogenesis in the North Atlantic (1948–2004). *Monthly Weather Review*, **136 (4)**, 1284–1304, doi:10.1175/2007MWR2245.1.

McTaggart-Cowan, R., T. J. Galarneau, L. F. Bosart, R. W. Moore, and O. Martius, 2013: A global climatology of baroclinically influenced tropical cyclogenesis. *Monthly Weather Review*, **141 (6)**, 1963–1989, doi:10.1175/MWR-D-12-00186.1.

Messner, J. W., G. J. Mayr, and A. Zeileis, 2017: Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Monthly Weather Review*, **145 (1)**, 137–147, doi:10.1175/MWR-D-16-0088.1.

Murakami, H., G. Villarini, G. A. Vecchi, W. Zhang, and R. Gudgel, 2016: Statistical–dynamical seasonal forecast of north atlantic and u.s. landfalling tropical cyclones using the high-resolution gfdl flor coupled model. *Monthly Weather Review*, **144**, 2101–2123, doi:10.1175/MWR-D-15-0308.1.

Murphy, A. H., 1971: A note on the ranked probability score. *Journal of Applied Meteorology and Climatology*, **10 (1)**, 155–156.

Murphy, A. H., 1973: A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, **12 (4)**, 595–600, doi:10.1175/1520-0450(1973)012⟨0595:ANVPOT⟩2.0.CO;2.

Murphy, A. H., 1977: The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review*, **105 (7)**, 803–816, doi:10.1175/1520-0493(1977)105⟨0803:TVOCCA⟩2.0.CO;2.

NOAA NCEI, 2022: NOAA National Centers for Environmental Information (NCEI) U.S. Billion-Dollar Weather and Climate Disasters. Accessed 14 December 2021. Available online at https://www.ncei.noaa.gov/access/billions/, doi:10.25921/stkw-7w73.

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. *Journal of 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. *Geophysical Research Letters*, **47**, e2020GL090 095, doi:10.1029/2020GL090095.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **126 (563)**, 649–667, doi:10.1002/qj.49712656313.

Richardson, D. S., 2003: Predictability and economic value. *Seminar on Predictability of weather and climate*, Shinfield Park, Reading, ECMWF, [Available online at https://www.ecmwf.int/sites/default/files/elibrary/2003/11922-predictability-and-economic-value.pdf.].

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.

Rogers, R., and Coauthors, 2006: The intensity forecasting experiment: A NOAA multiyear field program for improving tropical cyclone intensity forecasts. *Bulletin of the American Meteorological Society*, **87 (11)**, 1523–1538, doi:10.1175/BAMS-87-11-1523.

Rogers, R., and Coauthors, 2013: NOAA'S hurricane intensity forecasting experiment: A progress report. *Bulletin of the American Meteorological Society*, **94 (6)**, 859–882, doi:10.1175/BAMS-D-12-00089.1.

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., and D. Möller, 2015: Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *Ann. Appl. Stat.*, **9 (3)**, 1328–1349, doi:10.1214/15-AOAS843.

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. *Monthly Weather Review*, **148**, 3489–3506, doi:10.1175/MWR-D-20-0096.1.

Schraff, C., and R. Hess, 2021: A description of the nonhydrostatic regional COSMO-Model. Part III: Data assimilation. Consortium for Small-Scale Modelling, 102 pp., doi:10.5676/dwd_pub/nwv/cosmo-doc_6.00_III.

Schreck III, C. J., J. Molinari, and A. Aiyyer, 2012: A global view of equatorial waves and tropical cyclogenesis. *Monthly Weather Review*, **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. *Journal of the Atmospheric Sciences*, **68**, 195–209, doi:10.1175/2010JAS3396.1.

Schättler, U., and U. Blahak, 2021: A description of the nonhydrostatic regional COSMO-Model. Part V: Initial and boundary data for the COSMO-model. Consortium for Small-Scale Modelling, 84 pp., doi:10.5676/dwd_pub/nwv/cosmo-doc_6.00_V.

Shapiro, L. J., 1982: Hurricane climatic fluctuations. Part II: Relation to large-scale circulation. *Monthly Weather Review*, **110**, 1014–1023, doi:10.1175/1520-0493(1982)110⟨1014:HCFPIR⟩2.0.CO;2.

Simpson, R. H., 1974: The hurricane disaster potential scale. *Weatherwise*, **27 (8)**, 169.

Sobel, A. H., A. A. Wing, S. J. Camargo, C. M. Patricola, G. A. Vecchi, C.-Y. Lee, and M. K. Tippett, 2021: Tropical cyclone frequency. *Earth's Future*, e2021EF002275, doi:10.1029/2021EF002275.

Stephenson, D. B., C. A. Coelho, and I. T. Jolliffe, 2008: Two extra components in the Brier score decomposition. *Weather and Forecasting*, **23**, 752–757, doi:10.1175/2007WAF2006116.1.

Swets, J. A., 1988: Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293, doi:10.1126/science.3287615.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proceedings, ECMWF Workshop on Predictability*, ECMWF, 1-25, [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.].

Tang, B. H., and Coauthors, 2020: Recent advances in research on tropical cyclogenesis. *Tropical Cyclone Research and Review*, **9 (2)**, 87–105, doi:10.1016/j.tcrr.2020.04.004.

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

Tobin, J., 1958: Estimation of relationships for limited dependent variables. *Econometrica*, 24–36, doi:10.2307/1907382.

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.

Van Oldenborgh, G. J., and Coauthors, 2017: Attribution of extreme rainfall from Hurricane Harvey, August 2017. *Environmental Research Letters*, **12 (12)**, 124 009, doi:10.1088/1748-9326/aa9ef2.

Vannitsem, S., and Coauthors, 2021: Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World. *Bulletin of the American Meteorological Society*, **102 (3)**, E681–E699, doi:10.1175/BAMS-D-19-0308.1.

Villarini, G., B. Luitel, G. A. Vecchi, and J. Ghosh, 2019: Multi-model ensemble forecasting of north atlantic tropical cyclone activity. *Climate Dynamics*, **53 (12)**, 7461–7477, doi:10.1007/s00382-016-3369-z.

Villarini, G., and G. A. Vecchi, 2012: North Atlantic power dissipation index (PDI) and accumulated cyclone energy (ACE): Statistical modeling and sensitivity to sea surface temperature changes. *Journal of Climate*, **25 (2)**, 625–637, doi:10.1175/JCLI-D-11-00146.1.

Vitart, F., 2009: Impact of the Madden Julian Oscillation on tropical storms and risk of landfall in the ECMWF forecast system. *Geophysical Research Letters*, **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. *Monthly Weather Review*, **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 Bulletin*, **61**, 23–28.

Vitart, F., and T. N. Stockdale, 2001: Seasonal forecasting of tropical storms using coupled GCM integrations. *Monthly Weather Review*, **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. *Bulletin of the American Meteorological Society*, **98**, 163–173, doi:10.1175/BAMS-D-16-0017.1.

Vogel, P., P. Knippertz, A. H. Fink, A. Schlueter, and T. Gneiting, 2018: Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical africa. *Weather and Forecasting*, **33 (2)**, 369–388.

Walz, E.-M., M. Maranan, R. van der Linden, A. H. Fink, and P. Knippertz, 2021: An imerg-based optimal extended probabilistic climatology (epc) as a benchmark ensemble forecast for precipitation in the tropics and subtropics. *Weather and Forecasting*.

Wang, C., C. Deser, J.-Y. Yu, P. DiNezio, and A. Clement, 2017: *El Niño and Southern Oscillation (ENSO): A Review*, 85–106. Springer Netherlands, Dordrecht, doi:10.1007/978-94-017-7499-4_4.

Wang, Z., G. Zhang, T. J. Dunkerton, and F.-F. Jin, 2020: Summertime stationary waves integrate tropical and extratropical impacts on tropical cyclone activity. *Proceedings of the National Academy of Sciences*, **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. *Journal of the Atmospheric Sciences*, **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. *Monthly Weather Review*, **129**, 2677–2694, doi:10.1175/1520-0493(2001)129⟨2677:RTMAPO⟩2.0.CO;2.

Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Monthly Weather Review*, **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., Academic Press, 676 pp.

Zhai, A. R., and J. H. Jiang, 2014: Dependence of US hurricane economic loss on maximum wind speed and storm size. *Environmental Research Letters*, **9 (6)**, 064 019, doi:10.1088/1748-9326/9/6/064019.

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

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

Zhang, W., G. A. Vecchi, G. Villarini, H. Murakami, R. Gudgel, and X. Yang, 2017b: Statistical–dynamical seasonal forecast of western north pacific and east asia landfalling tropical cyclones using the gfdl flor coupled climate model. *Journal of Climate*, **30 (6)**, 2209–2232, doi:10.1175/JCLI-D-16-0487.1.

Acknowledgements

My greatest thanks go to God, my creator, who has once again proven Himself faithful even in the greatest difficulties. I ultimately owe the completion of this dissertation to his care and unconditional love.

I want to express my deepest gratitude to my supervisors Andreas H. Fink and Michael Riemer, and to our group leader and head of institute, Peter Knippertz, for their continuing support and advice, encouraging optimism, and patience. Thank you for the trust you have placed in me by creating the PhD position and for giving me the freedom to be involved in shaping the associated C3 project.

Scientific approaches and results benefit from many in-depth discussions, and in this regard my special thanks are due to Andreas Schlüter, Philipp Zschenderlein, Enrico Di Muzio, and Marlon Maranan for the meteorological aspects of my topic. I would like to particularly thank Eva-Maria Walz, Benedikt Schulz, and Christoph Fischer for their dedicated support in answering many questions about statistics, forecast verification, and computer science. Moreover, I am deeply grateful to Tilmann Gneiting, Sebastian Lerch, and Peter Vogel for their valuable advice on various aspects of statistical forecasting and model validation, which I received through regular meetings and personal communication. The work leading to this dissertation was embedded in the C3 project as part of the DFG-funded Transregional Collaborative Research Center SFB/TRR 165 ”Waves to Weather”. I greatly appreciated the opportunity to be a member of this consortium with its supportive program for early career scientists, which enabled me to attend several conferences and workshops and to visit research institutes in the USA. In this context, I thank Sharanya Majumdar, Philip Klotzbach, Tim Hewson, Ron McTaggart-Cowan, Paul Roundy, Lance Bosart, Eric Maloney, Ryan Torn, and others for comments that helped shape the work presented in this thesis.

In collaboration with Maurus Borne, I had the chance to apply the developed models in a real-time prediction mode, a side project where I could learn a lot. Many thanks to Gerhard Brückel, Jan-Hendrik Daub, and Robert Redl for help in setting up this real-time server and for their support with other technical issues over the past years. I also thank Friedericke Schönbein, Alexandra Beideck, and Doris Stenschke, who helped me with

various administrative matters. A very special thank you goes out to Ulrike Enderle, who took the time to proofread parts of the thesis.

My acknowledgements would not be complete without thanking my wonderful wife for her love, support, and encouragement. Finally, I want to thank my parents, brothers, and friends, who stood by my side and helped me in so many different ways. Without all of you, I could not have afforded to tackle this endeavor!