Forecasts of Hurricanes Using Large-Ensemble Outputs

JONATHAN LIN AND KERRY EMANUEL

Lorenz Center, Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts

JONATHAN L. VIGH

Joint Numerical Testbed Program, Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado

ABSTRACT

This paper describes the development of a model framework for Forecasts of Hurricanes Using Large-Ensemble Outputs (FHLO). FHLO quantifies the forecast uncertainty of a tropical cyclone (TC) by generating probabilistic forecasts of track, intensity, and wind speed that incorporate the state-dependent uncertainty in the large-scale field. The main goal is to provide useful probabilistic forecasts of wind at fixed points in space, but these require large ensembles [$O(1000)$] to flesh out the tails of the distributions. FHLO accomplishes this by using a computationally inexpensive framework, which consists of three components: 1) a track model that generates synthetic tracks from the TC tracks of an ensemble numerical weather prediction (NWP) model, 2) an intensity model that predicts the intensity along each synthetic track, and 3) a TC wind field model that estimates the time-varying two-dimensional surface wind field. The intensity and wind field of a TC evolve as though the TC were embedded in a time-evolving environmental field, which is derived from the forecast fields of ensemble NWP models. Each component of the framework is evaluated using 1000-member ensembles and four years (2015–18) of TC forecasts in the Atlantic and eastern Pacific basins. We show that the synthetic track algorithm generates tracks that are statistically similar to those of the underlying global ensemble models. We show that FHLO produces competitive intensity forecasts, especially when considering probabilistic verification statistics. We also demonstrate the reliability and accuracy of the probabilistic wind forecasts. Limitations of the model framework are also discussed.

1. Introduction

Tropical cyclones (TCs) are complex weather systems that bring flooding, storm surge, high winds, and other hazards to many coastal and island locations. Each year, TCs cause billions of dollars in damage to businesses and property and result in the loss of numerous lives (Pielke et al. 2008). To mitigate such losses and allow vulnerable populations to undertake life-saving preparations, TC forecasts must be provided with sufficient lead time and be reasonably accurate and reliable. Forecasts of TCs have traditionally been separated into two categories: 1) track forecasts, which predict the location of the TC center, and 2) intensity forecasts, which, in the Atlantic

basin, predict the 1-min maximum sustained surface wind anywhere in the storm (Landsea and Franklin 2013). The forecasting community has used such metrics to quantify errors in forecast models. Through substantial model improvements, better observations, and improved data assimilation methods, forecast skills for track and intensity have been steadily improving (DeMaria et al. 2014).

One of the advantages of separating TC forecasts into track and intensity is that it allows for straightforward evaluation of model performance. However, there are some significant drawbacks with this approach. First, by definition, deterministic forecasts do not quantify the uncertainty in the track and intensity of a TC, which can be an issue if the intensity of the TC strongly depends on its track. The concept of quantifying forecast uncertainty is not new. In 1992, the National Meteorological Center (NMC) and the European Centre for Medium-Range Weather Forecasts (ECMWF) began running

ensemble numerical weather prediction (NWP) systems to forecast the range of future weather conditions, by slightly perturbing the initial conditions and model physics of each member of the ensemble NWP system (Tracton and Kalnay 1993; Molteni et al. 1996). Since then, ensemble forecasts have proven to be vital to advancing probabilistic forecasts of weather, which in turn allow us to quantify weather-related risks (Gneiting and Raftery 2005).

The various components of the TC forecast (e.g., track, intensity, size, rainfall) affect the specific hazards (e.g., wind speed, waves, surge inundation, and riverine flood inundation) that are experienced at a given location. The specific impacts that then result depend on the vulnerability and exposure of the asset at that location. Thus, the ability to quantify the uncertainty of TC wind speed forecasts is of utmost importance, especially for vulnerable communities. Much work on the intrinsic predictability of TCs has focused on the importance of both large-scale influences (Emanuel et al. 2004; Zhang and Tao 2013; Kieu and Moon 2016) and convective-scale processes (Van Sang et al. 2008; Sippel and Zhang 2008; Judt et al. 2016). Initial condition uncertainty in the intensity and inner-core moisture has also been suggested to play a significant role in TC forecast uncertainty (Emanuel and Zhang 2016, 2017). In light of this, deterministic forecasts of TCs may be highly misleading, and probabilistic forecasts that can sample the inherent forecast uncertainty are preferred. It is no surprise, then, that ensemble prediction systems (EPSs) have enhanced probabilistic forecasts of TCs (Majumdar and Finocchio 2010; Hamill et al. 2011). Another approach to generating members of an ensemble is to combine the predictions of many different, independent, models, as is done for some "superensemble" models (Williford et al. 2003; Vijaya Kumar et al. 2003); in fact, even a simple average of the forecasted tracks from a variety of dynamical models has been shown to outperform any of the individual models (Goerss 2000).

Since the wind field of a TC can span hundreds of kilometers, the position and intensity of a TC are insufficient for specifying the conditions at fixed points. In practice, interested parties should be more concerned about the wind speed probability distribution at a single location (pointwise wind speed probabilities), rather than the TC's maximum wind speed or exact center location. Unfortunately, due to computational constraints, ensemble NWP models are typically run with inadequate horizontal resolution to resolve strong gradients in pressure and temperature that are commonly found in TCs, which often leads to large underestimation of the wind speeds in the simulated TCs (Gentry and Lackmann 2010; Gopalakrishnan et al. 2011). To make

matters worse, an ensemble NWP system usually does not have enough ensemble members to flesh out the highest wind speeds of a pointwise wind distribution. Inadequate horizontal resolution and small ensemble size mean that it is impractical to use the (insufficiently) resolved wind fields from ensemble NWP models to produce well-resolved probabilistic, pointwise forecasts of wind speeds.

To address these issues and communicate the pointwise uncertainty of TC winds, DeMaria et al. (2009) developed the Monte Carlo probability[1] (MCP) model to estimate the pointwise probability of surface winds exceeding the 34-, 50-, and 64-kt ($1\,\mathrm{kt} \approx 0.51\,\mathrm{m\,s^{-1}}$) wind speed thresholds. For each forecast, the MCP model generates computationally inexpensive track and intensity realizations by adding random errors to the National Hurricane Center's official forecasts for track, intensity, and wind radii. The errors are sampled from the official forecast track/intensity/wind radii error distributions over the most recent 5 years. However, this method forces the uncertainty in the model to exactly equal the observed average error; in practice, this usually does not reflect the state-dependent uncertainty for a given forecast. Since the forecast uncertainty for each TC can vary greatly, the climatological uncertainty may poorly reflect the true uncertainty in the track/intensity/wind radii forecasts. DeMaria et al. (2013) improved the MCP model to separate the climatological error distributions into three categories, based on the Goerss predicted consensus error (GPCE), a parameter that measures the extent of the track spread among an ensemble system (Goerss 2007). Goerss (2007) showed that the degree of uncertainty of a forecast could be coarsely predicted based on a low, medium, or high GPCE, and separated the climatological error bins accordingly. While the inclusion of GPCE improved the sharpness of the probabilistic wind forecasts, the MCP model still uses climatologically based forecast errors (stratified by basin) to generate uncertainty in the official forecast. Ensemble-based uncertainties, on the contrary, characterize the state-dependent uncertainty in the large-scale system. Therefore, a model that is able to draw upon the uncertainty represented in the ensemble members of an NWP model, while maintaining the computationally inexpensive components of the MCP model, could improve on probabilistic wind speed forecasts.

We aim to develop a computationally inexpensive pointwise TC wind speed prediction framework that is

---

[1] The National Hurricane Center's operational version of this model is called the Tropical Cyclone Wind Speed Probabilities Product.

capable of incorporating the state-dependent forecast uncertainty. To be practical, the framework should be capable of quickly generating the large number of ensemble members necessary to create a robust probabilistic forecast. Our proposed framework will address these concerns by using a large-ensemble prediction system in which the ensemble members are computationally cheap, stochastic realizations reflecting the uncertainty derived from dynamical ensemble models. Furthermore, our framework is designed to readily scale with advancements in the physics, resolution, and size of ensemble NWP models, especially since there is still much room for improvement in forecasting the intensity of TCs (Emanuel and Zhang 2016).

The paper is organized as follows. Section 2 describes the datasets used in this study. A detailed description of Forecasts of Hurricanes Using Large-Ensemble Outputs (FHLO) is provided in section 3, followed by an evaluation of the model framework in section 4. Section 5 concludes with a summary and discussion.

2. Data

In this study, we use forecast data from EPSs as the primary inputs to FHLO. We use two EPSs in particular: the 51-member global ensemble of ECMWF, and NOAA's 21-member Global Ensemble Forecast System (GEFS). Both ensembles are run multiple times per day and provide an estimate of forecast uncertainty. All ensemble forecast data are obtained from an online portal generously made available by the THORPEX Interactive Grand Global Ensemble (TIGGE) project (Bougeault et al. 2010; Swinbank et al. 2016). The data are obtained on a 0.5° × 0.5° resolution grid. The analyzed tracks of TCs in ensemble forecast models are obtained from the TIGGE Model Tropical Cyclone Track Dataset (National Centers for Environmental Prediction et al. 2008), which is obtained from the Research Data Archive (RDA) maintained by the National Center for Atmospheric Research (NCAR).

We also require initial conditions at each initialization time. For the ECMWF ensemble, we use the reanalysis fields from the ERA5 atmospheric reanalysis (Hersbach 2016; Copernicus Climate Change Service 2017). Note that these fields can have significant differences from the analysis fields of the operational ECMWF model. However, we require high-resolution soundings in order to estimate the potential intensity. Since the high-resolution analysis fields of the operational ECMWF model are not available to the public, we have used the ERA5 fields. For the GEFS, we use the analysis fields from the Global Forecast System (GFS) analysis fields at each model initialization time, obtained through the RDA at NCAR (National Centers for Environmental Prediction 2015). All analysis fields are obtained on a 0.25° × 0.25° resolution grid. As part of initialization, we also require initialization of the TC itself. To obtain real-time estimates of TC position, intensity, and wind structure, we obtain automated tropical cyclone forecasting (ATCF) a-decks for each model initialization time (Sampson and Schrader 2000).

To obtain real-time estimates of the sea surface temperature, we use the National Centers for Environmental Information's 0.25° Optimum Interpolation Sea Surface Temperature (OISST) dataset (Reynolds et al. 2008). The sea surface temperature data is provided daily.

Finally, we use the HURDAT2 best track data to evaluate the performance of the model (Landsea and Franklin 2013).

3. The large-ensemble model for tropical cyclones

In this work, we model TCs by assuming that a TC vortex is embedded within an evolving large-scale environmental field that ultimately determines the TC's intensity. From an ensemble NWP model, such as the GEFS, we estimate environmental quantities relevant to the intensity of a TC, such as the saturation entropy deficit and vertical wind shear (Emanuel et al. 2004). By deriving these quantities from multiple members of an ensemble model, we can sample the internal variability in each environmental field. The track module of the large-ensemble model generates realistic tracks from the set of ensemble TC tracks. These tracks, combined with the various environmental fields along these tracks, serve as input into FAST (Emanuel 2017), a fast TC intensity model that emulates an idealized, axisymmetric TC model (Emanuel et al. 2004). FAST evaluates the intensity, or maximum azimuthal wind speed, along a specified track through one realization of the large-scale environmental field. Finally, the intensity and environmental fields are used as inputs into a parametric surface wind model, to generate a full spatial wind field of the TC (Chavas et al. 2015). Figure 1 summarizes the overall flow of information in the large-ensemble model. Each component of the complete model framework is explained in depth in the following sections.

a. Synthetic track model

In this section, we describe a track algorithm that draws information from the ensemble track covariance to generate a large number of statistically indistinguishable synthetic tracks. The model is physically motivated by the beta-and-advection model (Marks 1992), which assumes that TCs are advected by some large-scale steering flow. Since we do not expect the large-scale
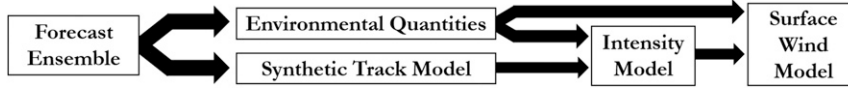
FIG. 1. Flow of information in FHLO.

flow to have considerable fluctuations on short (hourly) time scales, we expect there to be some correlation between the translational speed vector from one time step to the next. Note that the forecasted center of TCs is typically output in 3- or 6-h increments.

In light of this, we model the distribution of TC translational speeds using a Markov-chain solution. This means we condition on the previous time step translational speed, u_{t-1} and v_{t-1}, where u_{t-1} and v_{t-1} are the zonal and meridional translational speeds at time step $t-1$, respectively, to determine the translational speed at the next discrete time step. Mathematically, this corresponds to $P(u_t, v_t | u_{t-1}, v_{t-1})$. To properly describe this conditional distribution, we expand this as

$$P(u_t, v_t | u_{t-1}, v_{t-1}) = \frac{P(u_{t-1}, v_{t-1}, u_t, v_t)}{P(u_{t-1}, v_{t-1})}.$$

Next, we model both joint probability distributions as a mixture of k Gaussian distributions:

$$P(u_{t-1}, v_{t-1}, u_t, v_t) = \sum_{i=1}^{k} w_i N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad \text{and}$$

$$P(u_{t-1}, v_{t-1}) = \sum_{j=1}^{m} w_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j),$$

where w_i are the weights of each Gaussian mixture, such that $\sum_{i=1}^{k} = 1$. The quantities $\boldsymbol{\mu}_i$, $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_i$, and $\boldsymbol{\Sigma}_j$ must be estimated using the track displacements from the ensemble model. In this study, we set $k = 1$, though future work can explore how increasing k affects the track model.

Given (u_{t-1}, v_{t-1}), we can step the model forward by drawing from the conditional probability above. Integrating forwards in time will generate a synthetic track. The statistical algorithm is relatively fast, as 1000 synthetic tracks can be generated in approximately five minutes on a conventional laptop. For robustness, we require that at least 75% of the global ensemble member tracks have not dissipated before proceeding to the next time step.

Note that this method directly depends on the skill of the ensemble prediction system. While this means that the accuracy of the ensemble covariance algorithm should scale alongside the general accuracy of ensemble NWP models, it also indirectly ties the track algorithm

to how accurately the ensemble NWP models simulate the intensity of the analyzed TC. If the TC dissipates too early in the ensemble prediction model, the ensemble track covariance model will generate tracks that also dissipate too early.

Since a potentially unlimited number of synthetic tracks can be generated, robust probabilistic statistics can also be generated. Figure 2 shows an example of the corresponding 75-km strike probability for the forecast of Hurricane Irma (2017), initialized at 0000 UTC 5 September 2017. While we observe that the density of the 75-km strike probability corresponds well with the density of the actual ensemble tracks for this single case, a large sample size with probabilistic evaluation is needed to fully evaluate the quality of the framework's track predictions. This will be further examined in section 3d.

b. Intensity model

The intensity model evaluates the surface azimuthal wind speed along a particular track. Though any computationally inexpensive intensity model can be used in the large-ensemble framework, we choose to use the FAST system, a pair of coupled, nonlinear ordinary differential equations that describe the evolution of V, the maximum azimuthal wind, and m, an inner-core moisture variable that is bounded between 0 and 1 (Emanuel and Zhang 2017; Emanuel 2017). The choice was motivated primarily because the system is framed around physically based parameters that can be easily derived from ensemble fields. The equations are included below:

$$\frac{dV}{dt} = \frac{1}{2} \frac{C_k}{h} \left[\underbrace{\alpha \beta V_p^2 m^3}_{(1)} - \underbrace{(1 - \gamma m^3) V^2}_{(2)} \right], \tag{1}$$

$$\frac{dm}{dt} = \frac{1}{2} \frac{C_k}{h} \left[\underbrace{(1 - m) V}_{(3)} - \underbrace{\chi S m}_{(4)} \right], \tag{2}$$

$$\beta = 1 - \varepsilon - \kappa, \tag{3}$$

$$\gamma = \varepsilon + \alpha \kappa, \tag{4}$$

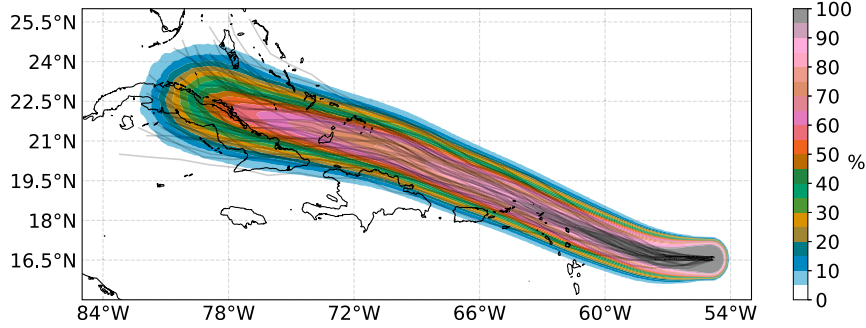$$\varepsilon = \frac{T_s - T_o}{T_s}, \quad \text{and} \tag{5}$$

FIG. 2. The 75-km strike probability, for Hurricane Irma in the Atlantic basin, using 1000 synthetic tracks generated from the ECMWF ensemble. Forecast initialized at 0000 UTC 5 Sep 2017. Overlaid lines depict TC centers from the ECMWF ensemble.

$$\kappa = \frac{\varepsilon}{2} \frac{C_k}{C_d} \frac{L_v}{R_d} \frac{q_s^*}{T_s}, \qquad (6)$$

where C_k is the surface enthalpy coefficient, C_d is the surface drag coefficient, h is the atmospheric boundary layer depth, V_p is the potential intensity, α is an ocean interaction parameter that varies between 0 and 1, χ is the midlevel saturation entropy deficit, S is the 250–850-hPa vertical wind shear, T_s is the surface temperature, T_o is the outflow temperature, L_v is a constant latent heat of vaporization, R_d is the dry gas constant, q_s^* is the surface saturation specific humidity, and ε is the thermodynamic efficiency [see Emanuel (2017) for further details]. Note that in Emanuel (2017), $\chi = 2.2$, which is a typical value for the saturation entropy deficit, but as discussed presently, we will evaluate its time-evolving value from the global ensemble. The behavior of this system is controlled by four key terms: 1) a spinup term that represents intensification of the vortex toward its potential intensity because of surface fluxes, 2) a spindown term from the thermodynamic dampening influence of downdrafts in the inner core, 3) a moistening term that represents surface moisture fluxes from the ocean, and 4) a drying term that mimics eddy entropy fluxes into the TC's eyewall. Though remarkably simple, the FAST equations do have some limitations, which are discussed in the conclusion.

The FAST equations are integrated forwards in time using a Runge–Kutta fourth-order numerical scheme with a time step of 450 s. As in Emanuel (2017), we set $C_k = 1.2 \times 10^{-3}$, $h = 1400$ m, $\varepsilon = 0.33$, $\kappa = 0.1$. Thus, given an initial intensity and inner-core moisture, as well as the vertical wind shear, saturation entropy deficit, and potential intensity along a track, we can solve for the time evolution of V and m. It is worth nothing that in order to properly calculate α, the ocean interaction parameter that modulates V_p, one needs the ocean mixed layer depth h_m as well as the sub–mixed layer thermal

stratification Γ. For simplicity, we choose to use climatological values for h_m and Γ (Levitus 1982). Future work will incorporate this information from a real-time ocean forecast model. Ocean mixing is switched off whenever the sub–mixed layer depth is larger than the depth of the ocean. Finally, a 0.25° bathymetric dataset is used to determine when the center of the TC is over land, during which the potential intensity is set to zero.

In this framework, the three most important environmental quantities that influence the intensity of a TC are 1) vertical wind shear, 2) saturation entropy deficit, and 3) potential intensity. These dynamic and thermodynamic quantities are defined to be environmental fields, and thus should be evaluated assuming that the considered TC does not exist. To progress with the perspective of modeling a TC in a synthetic environment, we need to remove any effects on these fields that are induced by an analyzed TC.

c. Environmental quantities

1) VERTICAL WIND SHEAR

To calculate the environmental vertical wind shear, the circulation induced by the TC must be removed. One method, for instance, averages winds over some distance larger than the radius of the inner core of the TC (DeMaria and Kaplan 1994), while other methods set the vorticity and divergence to zero within a specified distance from the TC center, and invert the Poisson equation to find the streamfunction and velocity potential associated with the vortex (Davis et al. 2008; Galarneau and Davis 2013). Since we desire a continuous spatial field of environmental winds $\mathbf{u}_{\text{env}}$, we choose the latter method, setting the relative vorticity and divergence of the environmental field to zero within an inversion radius of r^* from the vortex center. We use $r^* = 400$ km, where the magnitude of the axisymmetric component of the vortex's relative vorticity becomes comparable to that of the

environmental field. Note that $r^* = 400$ km is around the middle range of r^* values used by Galarneau and Davis (2013), and close to the median outer radius of TCs, inferred using scatterometer data (Chavas and Emanuel 2010). Fixing r^* is perhaps not the best choice, as Galarneau and Davis (2013) showed that the resulting environmental winds are in fact sensitive to r^*. Future work could include stochastic perturbations to r^*, or optimizing r^* from environmental profiles of relative vorticity. Defining ζ_{vortex} and δ_{vortex} as the vorticity and divergence identified as part of the vortex, respectively, we have

$$
\begin{aligned}
\nabla^2 \psi_{\text{vortex}} &= \zeta_{\text{vortex}} \\
\nabla^2 \Phi_{\text{vortex}} &= \delta_{\text{vortex}} \\
\mathbf{u}_{\text{vortex}} &= \nabla \Phi_{\text{vortex}} + \hat{\mathbf{k}} \times \nabla \psi_{\text{vortex}} \\
\mathbf{u} &= \mathbf{u}_{\text{env}} + \mathbf{u}_{\text{vortex}} ,
\end{aligned} \tag{7}
$$

where $\mathbf{u}$ is the wind velocity vector, ψ is the streamfunction, Φ is the velocity potential, ∇ is the gradient operator, and ∇^2 is the Laplacian operator. To solve for $\mathbf{u}_{\text{vortex}}$, we must invert the Laplacian operator on a sphere to obtain the corresponding streamfunction and velocity potential. Details on this inversion are included in the appendix. After calculation of the environmental wind at 250- and 850-hPa, we subtract the two to obtain the environmental vertical wind shear. We chose these two levels based on DeMaria and Kaplan (1994), who found that the vertical wind shear between the 250- and 850-hPa levels correlates well with intensity changes in tropical cyclones, though alternative pressure levels could also be used to calculate the vertical wind shear.

2) MIDLEVEL VENTILATION

Midlevel ventilation, or the entrainment of low-entropy environmental air into a TC at midlevels, is one pathway by which vertical wind shear can interact with a TC and cause it to weaken (Simpson and Riehl 1958; Tang and Emanuel 2010). In the ventilation hypothesis, vertical wind shear leads to asymmetric processes that induce eddy fluxes and mixing between the TC eyewall and its environment. If the environmental air is sufficiently low in entropy, downdrafts will occur in the eyewall and disrupt warming of the inner core. For a TC with a saturated inner core, the normalized eddy fluxes that result from such ventilation are proportional to χS [see the appendix of Tang and Emanuel (2012)], where S is the 250–850-hPa vertical wind shear, and χ is a scalar that represents the saturation entropy deficit normalized by the air–sea thermodynamic disequilibrium, as in Eq. (8):

$$
\chi = \frac{s_m^* - s_{\text{env}}}{s_s^* - s_b}. \tag{8}
$$

The pseudoadiabatic entropy s can be approximated following Bryan (2008):

$$
s = c_p \log\left(\frac{T}{T_0}\right) - R_d \log\left(\frac{p_d}{p_0}\right) + \frac{L_v q}{T} - R_v q \log(\mathscr{H}),
$$

where c_p is the specific heat of dry at constant pressure, q is the specific humidity, $\mathscr{H}$ is the relative humidity, p_d is the dry pressure, R_v is the water vapor gas constant, s_m^* is the inner-core saturation entropy, s_{env} is the environmental entropy, s_s^* is the saturation entropy at the sea surface, and s_b is the entropy at the boundary layer. To calculate χ at a fixed pressure level p from gridded data, we first assume that temperature perturbations on pressure surfaces are small, and that the inner core is saturated (Emanuel et al. 2008), such that the numerator becomes

$$
s_p^* - s_p \approx \left[\frac{L_v q^*}{T}(1 - \mathscr{H}) + R_v q \log \mathscr{H} \right]\bigg|_p .
$$

To evaluate these quantities, we assume that the air at the sea surface is saturated and at the same temperature and pressure as the sea surface. We also assume that $s_b = s_{\text{LCL}}^* = s_p^*$, where s_{LCL}^* is the saturation entropy at the lifted condensation level. The first step assumes adiabatic motion from the boundary layer to the lifted condensation level, which is defined as the top of the boundary layer, and the last step follows from moist convective neutrality. Then, we have

$$
\chi_p = \frac{\left[L_v q^* T^{-1}(1 - \mathscr{H}) + R_v q \log \mathscr{H} \right]\big|_p}{s_s^* - s_p^*}.
$$

While the environmental saturation deficit is typically evaluated at $p = 600$ hPa (Emanuel 2013), consistent with the midlevel ventilation hypothesis, the 600-hPa level is not available through the TIGGE database. Instead, we calculate χ_p at $p = 500$ hPa and $p = 700$ hPa and take the gridpoint maximum between the two levels. To obtain the χ used in FAST, we take the Nth percentile of the distribution of the saturation entropy deficit within r_{env} of the TC center. Since any downdrafts that occur near the core are detrimental to the TC, we take relatively large values of N. The saturation entropy deficit typically increases away from the core, since deep convection near the inner core saturates the midlevels, such that if N is large enough, we are effectively diagnosing an environmental entropy deficit. To calculate the denominator, we take the median of the air–sea thermodynamic disequilibrium over the inner 200 km from the TC center. When the air–sea disequilibrium is negative, which can occur, for instance, at cold

SSTs, we set $\chi = \chi_d$. We also cap χ to a value of χ_d. We estimated the optimal values of N, r_{env}, and χ_d, by finding values that minimized the mean absolute error in intensity. We found $N = $ 90th percentile and $r_{env} = 1000$ km for the Atlantic basin, $N = $ 50th percentile and $r_{env} = 900$ km for the eastern Pacific basin, and $\chi_d = 4$. The large differences in N between the two basins is largely a result of differences in climatology; the density of tropical cyclone forecasts maximizes in thermodynamically favorable environments in the Atlantic, while the opposite is true in the eastern Pacific.

3) POTENTIAL INTENSITY

The potential intensity of a TC is a theoretical upper bound on its maximum wind speed (Emanuel 1986). Potential intensity has been verified as a reasonable upper bound on the intensity of real TCs (Emanuel 2000), and is defined as

$$V_p^2 = \frac{T_s - T_o}{T_o} \frac{C_k}{C_d} (k^* - k),$$

where k is the enthalpy of the boundary layer air, and k^* is the saturation enthalpy of the sea surface. To calculate potential intensity, an environmental sounding with a high resolution in the vertical is required. Since potential intensity must be evaluated at the eyewall of a TC, we must remove the effect of the global ensemble system's warm-core anomaly, which acts to reduce the buoyancy of a parcel lifted from the surface. Previous methods that attempt to account for this deficiency use time-lagged potential intensity fields (Emanuel et al. 2004). One weakness of this method is that it ignores any short-term variability of potential intensity. Instead, we opt to smooth out the effects of an analyzed TC. To remove the thermodynamic effect of an analyzed TC, we apply a 9-box smoother across a $10° \times 10°$ grid box centered on the TC, to the temperature, specific humidity, and sea level pressure fields, holding the boundary fixed and smoothing inwards. This first-order approximation successfully removes the potential intensity minimum near the vortex center and allows us to obtain a robust estimate of the environmental potential intensity. In this study, we calculate the spatially varying potential intensity using the analysis (initialization) fields, which have a high vertical resolution. The potential intensity field is then kept fixed throughout the forecast, though the potential intensity of the TC can still change as it moves in space.

d. Wind model

Since FAST outputs the maximum surface azimuthal wind speed, we use a parametric wind model and another parameterization to obtain the full TC wind field as well as the maximum surface wind speed. We first obtain the axisymmetric wind profile by using the physically based wind model developed by Chavas et al. (2015). We chose this model because of its basis on physical principles, though other wind models, such as the radii-CLIPER model (Knaff et al. 2007), could be used as well. The wind model developed in Chavas et al. (2015) separates the axisymmetric wind field into two regions, a convecting inner region, and a subsiding outer region, and the equations are below for convenience:

$$\frac{M_{inner}}{M_m} = \frac{2(r/r_m)^2}{1 + (r/r_m)^2}, \quad \text{and}$$

$$\frac{\partial M_{outer}}{\partial r} = \frac{2C_d}{W_{cool}} \frac{(rV)^2}{r_0^2 - r^2},$$

where $M = rV + (1/2)fr^2$ is the angular momentum per unit mass, M_{inner} (M_{outer}) is the angular momentum in the inner (outer) region, r is the radius from the vortex center, V is the azimuthal wind, r_0 is the outer radius (or radius of vanishing wind), r_m is the radius of maximum winds, W_{cool} is the subsidence rate from tropospheric radiative cooling, and $M_m(r_m, V_m)$ is the angular momentum at the radius of maximum wind, where V_m is the maximum azimuthal wind speed. Because we lack observations of W_{cool}, we set $2C_d/W_{cool} = 1\,\mathrm{s\,m^{-1}}$, for simplicity, though the model does have some dependence on the strength of W_{cool} (Chavas and Lin 2016). The full axisymmetric wind field can be resolved by fixing two of the three free parameters, r_m, r_0, and V_m. The maximum azimuthal wind speed V_m is readily obtained from FAST, and we choose to specify r_0 (details on selecting r_0 are described in section 3e). After obtaining the surface axisymmetric wind field, we apply a second model to obtain the asymmetric component of the wind field. This model is based on the isallobaric wind, which occurs whenever the vortex propagates with respect to the low-level wind, which will happen when there is vertical wind shear $\mathbf{S}$. In a reference frame moving with the low-level vortex, vorticity must be increasing downshear of the vortex center, and for this to happen, there must be low-level convergence. We crudely take this into account by representing the vorticity in terms of the maximum azimuthal wind $\mathbf{V}$ divided by a vortex radial length scale, and the vortex-relative vorticity advection by the shear vector times this vorticity divided by the same length scale. This yields the low-level convergence downshear of the low-level vortex. The associated convergent velocity component is then obtained by integrating over the same length scale. With some empirical adjusting of constants, this results in

$$\mathbf{u}_{\mathrm{net}} = \mathbf{V} + G\mathbf{u}_t + 0.1\mathbf{S}\frac{|\mathbf{V}|}{15}, \quad \text{and}$$

$$G = \min\left\{1, 0.8 + 0.35\left[1 + \tanh\left(\frac{\phi - 35}{10}\right)\right]\right\},$$

where $\mathbf{V}$ is the axisymmetric wind, $\mathbf{u}_t$ is the translational speed of the vortex, $\mathbf{S}$ is the 250–850-hPa vertical wind shear (m s^{-1}), ϕ is the latitude of storm center, $|\mathbf{V}|$ is the magnitude of the axisymmetric wind (kt), and $\mathbf{u}_{\mathrm{net}}$ is the net wind at the surface. Finally, the maximum surface wind speed is determined by taking the maximum magnitude of $\mathbf{u}_{\mathrm{net}}$ over the domain.

e. Initialization and parameter estimation

The final component of FAST pertains to initialization. A poor initialization can lead to significant errors in the short-range forecast period (Emanuel and Zhang 2016, 2017). Since we want to include initial intensity uncertainty in the probabilistic system, we create a synthetic perturbation of the TC intensity analysis over the past 24 h of the analyzed TC (see appendix for details). Uncertainties in the intensity observations are taken from climatological errors (Landsea and Franklin 2013). For each synthetic ensemble member, we use the environmental parameters from the analysis fields to drive FAST, and add a time-varying forcing term to the azimuthal wind [Eq. (1)] such that the modeled maximum surface wind best matches the synthetic perturbation of the observed TC intensity. The initialization period runs from 48 h before the initial forecast time until the initial forecast time, where the forcing term to the azimuthal wind equation then decays in magnitude as $\exp[-(t/t_0)^2]$, where t is the forecast lead time, and $t_0 = 1$ day.

To initialize the wind field, we take the initial analysis of the maximum extent of the 34-, 50-, and 64-kt winds in each quadrant [obtained from the combined automated response to query (CARQ) lines of the ATCF a-decks], and find the corresponding value of r_0 that allows the modeled asymmetric wind field to best match the analysis. Furthermore, to combat large negative biases in the axisymmetric wind model at radii where $3 \lessapprox r/r_m \lessapprox 6$ (see Fig. 9 in Chavas et al. 2015), we add a shape parameter k to the axisymmetric wind profile, that is, for $r > r_m$, $V(r, k) = V(r)^k$. Then, to initialize the wind field, we find r_0 and k such that the full asymmetric wind field best matches the analysis radii in each quadrant. Finally, if an official forecast of the radii in each quadrant is available, the optimal r_0 and k, given the official forecast of intensity, are used to interpolate r_0 and k forward in the forecast. Otherwise, r_0 and k are kept constant in the forecast. If an initial

analysis and forecast of the wind radii do not exist, we set $r_0 = 700$ km and $k = 1$. This may not be realistic but could be improved upon by using the model developed in Knaff et al. (2017), which predicts the maximum extent of the 34-, 50-, and 64-kt winds. We do not explicitly perturb r_0, though this could be the subject of future work.

4. Evaluation of the large-ensemble model

To robustly evaluate the skill of FHLO, we run 1000-member ensemble reforecasts for all 0000 and 1200 UTC cycle TC forecast cases in the Atlantic and eastern Pacific basins during the years 2015–18 (the choice of 1000 members is described in the following section). Since all of the aforementioned data used to generate a probabilistic forecast are available in real time, these reforecasts can be considered equivalent to late-cycle real-time forecasts. Since the skill of FHLO also depends on the skill of the ensemble, we run two variations of 1000-member ensembles, one using data from the ECMWF ensemble (FHLO-ECMWF) and the other from the GEFS (FHLO-GEFS). We also combine the two into a 2000-member superensemble, which we will denote FHLO-Super. We evaluate the performance of FHLO by using the HURDAT2 best track data as the observed track, intensity, and wind radii of each TC (Landsea and Franklin 2013). For each forecast case, we predict the track distribution, intensity distribution, and probability of exceedance for the 34-, 50-, and 64-kt wind speed thresholds.

a. Ensemble size

The choice of running 1000-member ensembles is motivated primarily by looking at probability distributions of wind speeds from TCs at a fixed point. To illustrate this, we generate pointwise forecasts of wind speed from Hurricane Maria at San Juan, Puerto Rico, using the ECMWF ensemble initialized on 0000 UTC 18 September 2017 as input into FHLO. For demonstration, we use 100- and 1000-member FHLO-ECMWF ensembles, as well as a 51-member FHLO-ECMWF model that only uses the tracks from the original ECMWF ensemble (RAW-ECMWF). Figure 3 shows the time-varying maximum wind speed from Hurricane Maria at San Juan, Puerto Rico, from the 1000-member FHLO-ECMWF and the 51-member RAW-ECMWF. From the eye test, there is a sampling issue with RAW-ECMWF; there simply are not enough ensemble members to resolve the tail of the distribution. To be more quantitative, we estimate the nondimensional damage ([0, 1]), f, that represents the fraction of property lost [see Eq. (1) of Emanuel (2011)]:
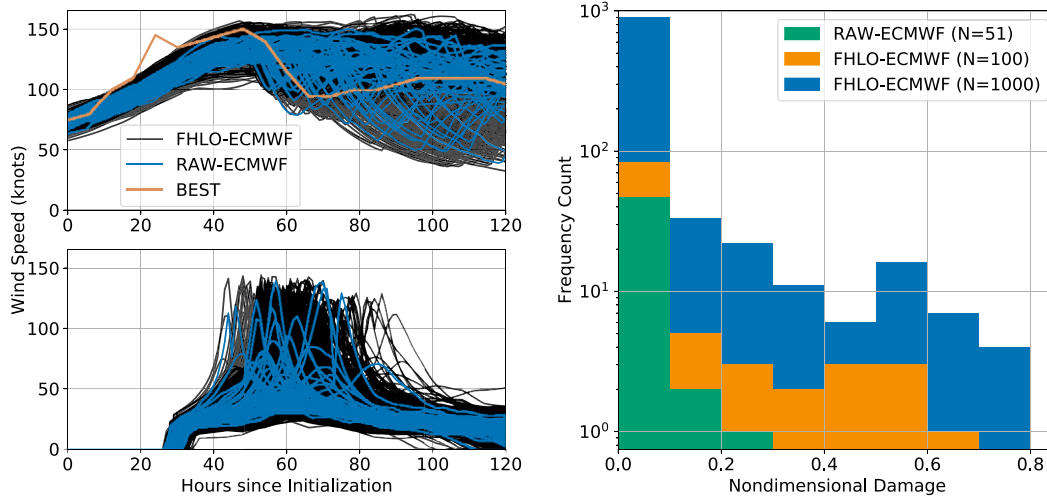
FIG. 3. Forecasts, initialized at 0000 UTC 18 Sep 2017, of the time-varying maximum wind speed from Hurricane Maria, (top left) across the entire storm and (bottom left) at San Juan, Puerto Rico. Forecasts use the original 1000-member FHLO-ECMWF model and a 51-member FHLO-ECMWF model (RAW-ECMWF) that uses only the original ECMWF ensemble tracks. BEST indicates the best track intensity of Hurricane Maria. (right) Distribution of nondimensional damage f, following Emanuel (2011), for wind speeds observed at San Juan at $t = 60$ h from initialization, for the RAW-ECMWF model (51 tracks), as well as the 100-member and 1000-member FHLO-ECMWF ensembles.

$$f = \frac{v_n^3}{1 + v_n^3}, \quad \text{and}$$

$$v_n \equiv \frac{\text{MAX}(V - V_{\text{thresh}}, 0)}{V_{\text{half}} - V_{\text{thresh}}},$$

where V is the maximum wind speed, V_{thresh} is the wind speed at which no damage occurs, and V_{half} is the wind speed at which half damage occurs. As the right column of Fig. 3 shows, there is an inferred probability of zero for $f > 0.3$ from the 51-member RAW-ECMWF. However, as we increase the size of the ensemble, the right tail of the distribution, which represents the most destructive scenarios, is better resolved. Thus, large-ensemble forecasts are necessary to flesh out the tail of wind distributions, which is often critical to decision making. With only a small number of ensemble members, it is extremely difficult to create smooth PDFs of pointwise wind forecasts. We settled on a 1000-member ensemble, since the marginal return on resolving the tail diminishes with further increases in ensemble size.

b. Track forecasts

To verify that the synthetic track algorithm produces tracks that are statistically similar to the set of TC tracks of a given EPS, we evaluate the track error distribution and spread from forecasts of 1) the mean of the synthetic tracks, and 2) the mean of the ensemble system. A two-sample Kolmogorov–Smirnov test indicates that the track error distributions of forecasts from both methods are statistically similar at all lead times, regardless of the ensemble, with $\alpha = 0.05$ (not shown).

Figure 4 shows the error between the observed track and the mean of the 1000-member ensemble, which is driven by the tracks of different EPSs. FHLO-GEFS has the smallest initial error for both basins, since the GEFS ensemble relocates the analyzed vortex to the best-guess position at initialization (Liu et al. 2000). In the Atlantic basin, FHLO-ECMWF outperforms all of the other individual models for all other lead times, while FHLO-GEFS has the best performance in the eastern Pacific. The multimodel superensemble, has the lowest mean absolute error (MAE) for almost all lead times in both basins, which is a well-known property of superensembles (Williford et al. 2003; Vijaya Kumar et al. 2003). Of course, the best metrics for the evaluation of a large ensemble are probabilistic metrics. A reliability diagram assesses the observed probabilities as a function of the forecasted probabilities. A model is said to be reliable if the observed frequencies of an event match the observed frequencies of its forecast probability—that is, the model performs well all the time without its forecasts being overconfident or underconfident. A perfectly reliable model falls along the 1:1 line in a reliability diagram. Figure 5 shows the reliability curve for the probability that the center of the TC is within 75-km of a particular grid point (using a 0.1° resolution grid), cumulative over 5 days. In general, the results demonstrate the reliability
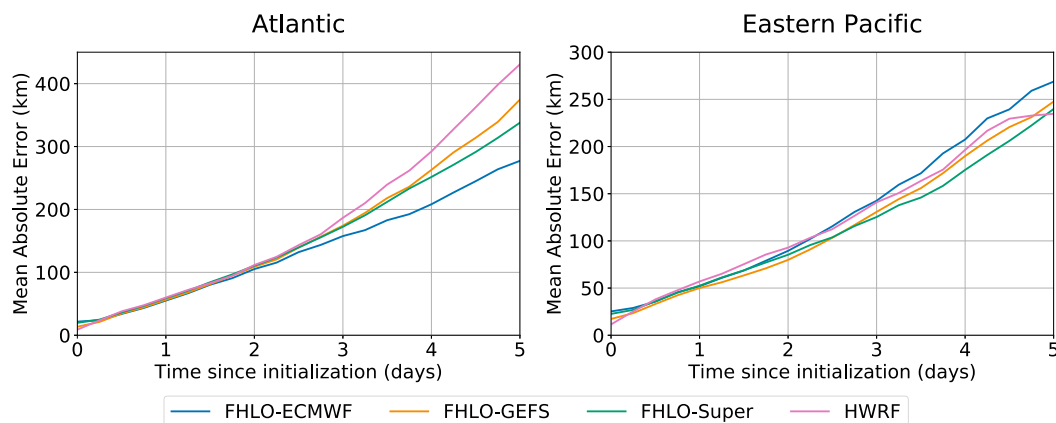
FIG. 4. Mean absolute error (km) in track as a function of the time since initialization, for the mean of 1000 synthetic tracks for the FHLO-ECMWF, FHLO-GEFS, and FHLO-Super ensembles, as well as the HWRF Model. Sample set includes forecasts from 2015 to 2018 in the (left) Atlantic basin and (right) eastern Pacific basin.

of the FHLO-ECMWF strike probabilities, as well as the overconfidence of the FHLO-GEFS strike probabilities. These results are similar to those obtained in Titley et al. (2020). We forgo a detailed evaluation of the synthetic tracks in lieu of evaluating the wind speed exceedance probabilities in section 4d, as the latter cannot be accurate if the former is not.

c. Intensity forecasts

In this section we evaluate the intensity forecasts of FHLO using traditional, deterministic statistics, as well as probabilistic metrics. When evaluating, we do not explicitly terminate a TC when its simulated intensity falls below the tropical depression threshold ($17\,\mathrm{m\,s^{-1}}$). Figure 6 shows a model comparison of the mean absolute error in intensity as a function of time since initialization, for forecasts where the initial intensity is greater than $30\,\mathrm{m\,s^{-1}}$ (the reason for applying this filter is described below). Since the mean absolute error is a deterministic statistic, we take the mean of all the members of each FHLO ensemble as its deterministic model forecast. In both basins, the intensity error is largely the same between FHLO-ECMWF, FHLO-GEFS, and FHLO-Super for the first two days, likely since the large-scale environments of the ECMWF ensemble and GEFS do not diverge significantly for these short-term time scales. The intensity errors begin to diverge around two days, after which we observe that FHLO-ECMWF and FHLO-Super outperform FHLO-GEFS. The FHLO-based intensity forecasts also have a larger absolute error than the Hurricane Weather Research and Forecasting (HWRF) Model, though the FHLO-based ensembles perform comparably to HWRF in the eastern Pacific. The FHLO-based ensembles have a slight negative bias

in the first two days of initialization. In general, however, the bias over the sample set is comparable to that of HWRF.

We also construct an idealized lower bound on the intensity error of the FHLO-based ensembles, to understand the best-case performance of a simple intensity model such as FAST. To gain insight into this, we take the exact same forecast cases, but instead use near-perfect initial conditions and analysis fields to derive the environmental parameters. The near-perfect initial conditions are achieved by running the aforementioned initialization procedure using the best track of the TC, forcing the model to the observed intensity. Once the
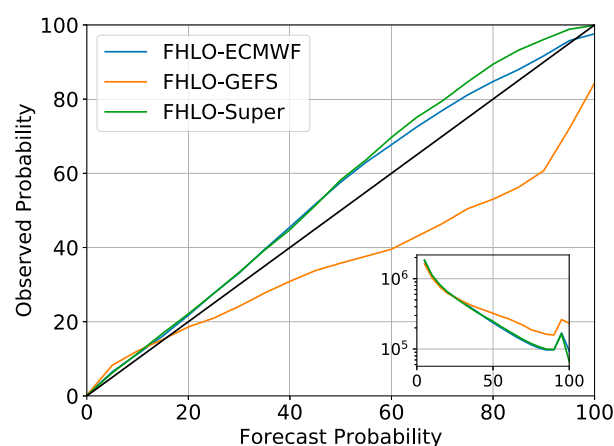

FIG. 5. The reliability curve for 75-km strike probability, cumulative over 5 days, using the FHLO-ECMWF, FHLO-GEFS, and FHLO-Super ensembles. Solid black line indicates perfect reliability. Sample set includes forecasts from 2015 to 2018 in the Atlantic and eastern Pacific basins, and sample size for each model, as a function of forecast probability, is shown in the inset.
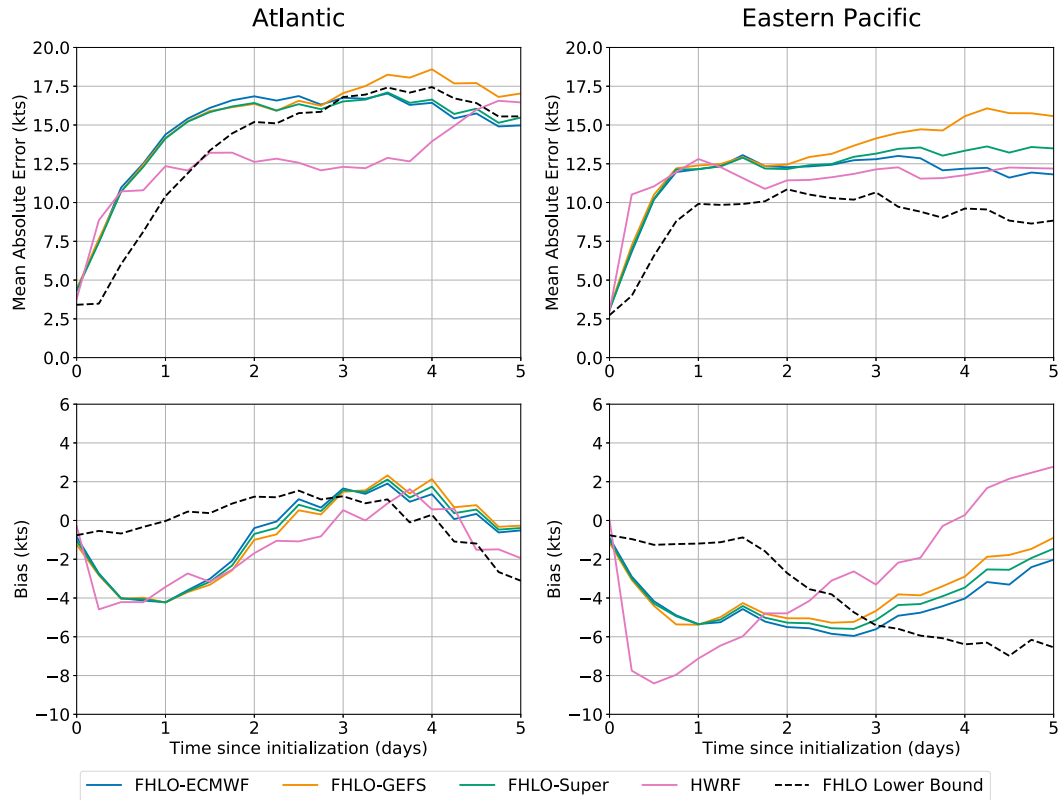
FIG. 6. The (top) mean absolute error and (bottom) bias of the FHLO-ECMWF, FHLO-GEFS, and FHLO-Super ensemble forecasts. Only forecasts where the initial intensity is greater than 30 m s^{-1} are included in the samples. The dashed black line represents a lower bound on the intensity error, achieved by using near-perfect initialization and a perfect track. Sample set includes forecasts from 2015 to 2018 in the (left) Atlantic basin and (right) eastern Pacific basin.

initialization procedure hits the initialization time, the model is allowed to evolve freely. Then, the lower-bound error is computed by the divergence of the FAST model from the best track intensity. This is shown as the dashed black curve in Fig. 6. In the eastern Pacific basin, the lower-bound curve suggests that with improved ensemble forecasts and model initialization, the intensity error of the FAST model will also decrease. This is somewhat less true in the Atlantic basin; the lower-bound curve suggests that better initialization could result in improvements in forecasts with lead times of up to 2 days. For lead times longer than 2 days, however, the lower-bound curve is nearly indistinguishable from the forecast errors. This may be due to low inherent predictability and/or model error stemming from suboptimal parameterization of the ventilation, though further investigation is outside the scope of this study. Note that these bounds are derived using the current framework of the model; improvements to the model itself could lower the bound further. Last, in some sense, this curve can be loosely compared to the model error of FAST. While simple models, such as FAST, are unable to physically resolve the atmosphere

and are often less accurate than more complex models, they are computationally inexpensive and can be used in large-ensemble studies such as this one. We believe that FAST makes a relatively good trade-off between simplicity and accuracy, though it is unlikely to be the best we can do.

Next, we evaluate the FHLO-based intensity forecasts using probabilistic metrics. A convenient metric to compare deterministic and probabilistic forecasts is the continuous ranked probability score (CRPS), which is the integrated squared difference between the cumulative distribution function (CDF) of the forecast and the observation:

$$\text{CRPS}(F, v_{\text{obs}}) = \int_{-\infty}^{\infty} [F(v) - \mathbb{1}(v - v_{\text{obs}})]^2 \, dv,$$

where $F(v)$ is the CDF of the forecast, v_{obs} is the observed intensity, and $\mathbb{1}$ is the Heaviside step function. For a deterministic forecast, the CRPS simplifies to the mean absolute error. Figure 7 shows the CRPS as a function of time since initialization, and we see similar
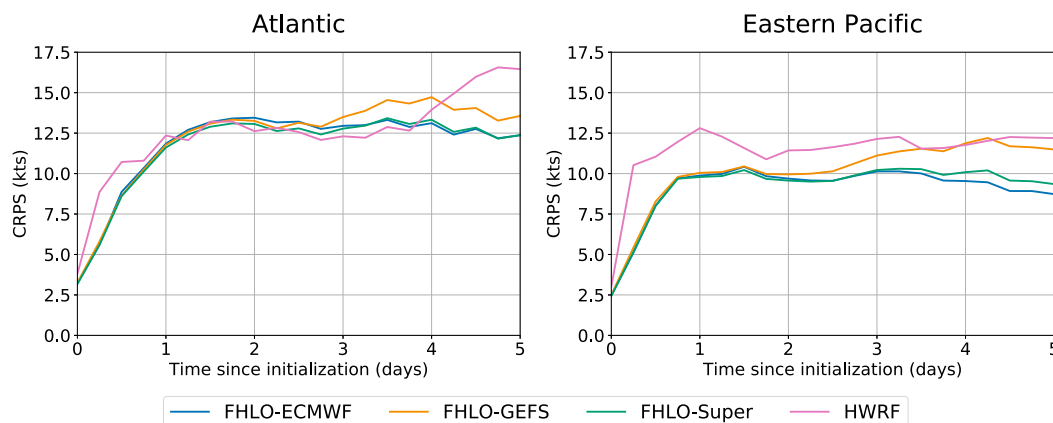
FIG. 7. The continuous ranked probability score of the FHLO-ECMWF, FHLO-GEFS, and FHLO-Super ensemble forecasts. Only forecasts where the initial intensity is greater than 30 m s^{-1} are included in the samples. Sample set includes forecasts from 2015 to 2018 in the (left) Atlantic basin and (right) eastern Pacific basin.

patterns in the results as compared to those of the mean absolute error. In the Atlantic basin, the CRPS of the FHLO-based ensembles is close to that of the deterministic HWRF forecast, for forecast lead times shorter than four days. For lead times longer than four days, the CRPS of the FHLO-based ensembles is lower than that of HWRF. In the eastern Pacific, the CRPS of the FHLO-based ensembles is lower than that of HWRF. This suggests that characterizing both the mean state and spread of the large-scale flow is paramount to quantifying the uncertainty in the future intensity of a TC.

Finally, it is worth commenting on the filter we used above, where we only considered forecasts where the initial intensity is greater than 30 m s^{-1}. The filter was specified in order to bypass the issue of cyclogenesis. In general, the mean absolute error in intensity increases as the initial intensity threshold in the evaluation filter is decreased (not shown), for a variety of reasons. While this may point to deficiencies in applying models based on idealized, axisymmetric TC theory to weak disturbances, there may be additional reasons for this behavior. For instance, a model that does not properly capture the observed intensification rate distribution but is tuned to provide sluggish intensification for weak storms could have the smallest mean absolute error in intensity, since most weak storms do not intensify into strong TCs. While the FAST system does capture the observed intensification rate distribution (Emanuel 2017), it has the tendency to overintensify weak disturbances, so long as the midlevel ventilation is not large. Regardless, much work has suggested that the intrinsic predictability of cyclogenesis and subsequent intensification is low (Sippel and Zhang 2008;

Zhang and Sippel 2009; Zhang and Tao 2013), and thus TC genesis remains a significant forecasting challenge (Rappaport et al. 2009).

d. Probabilistic wind speeds

In this section, we evaluate the 34-, 50-, and 64-kt wind speed exceedance probabilities that are generated from the FHLO-based ensembles. For each ensemble member, and at each time step, the full wind field of the TC is estimated. The evaluation procedure is similar to that in DeMaria et al. (2009), where the probabilities are generated by summing the number of ensemble members where the winds exceed a particular threshold, for each individual grid point. Following the evaluation procedure of DeMaria et al. (2009), we do not include forecast cases of extratropical transition. To do this, we truncate any forecasts that extend beyond when the official forecast predicts a transition to an extratropical storm. The wind probabilities should also account for the timing of dissipation, and thus we do not modify the forecast if the analyzed storm in the best track terminates before the end of the forecast, and vice versa. The latter situation is often more problematic for this model. This is because the ensemble tracks depend on how well the ensemble models resolve the TC on their relatively coarse grids. Finally, in this section, we only evaluate forecasts where the initial TC position is equatorward of 30°. This is a crude way to filter for storms that are characteristically more tropical, which is when we expect the FHLO-based ensembles to work the best. The aforementioned wind model also has more difficulty representing highly asymmetric wind fields, which is more likely to be the case for storms that are influenced by baroclinicity. Nevertheless, a large fraction (70% in

Atlantic, 99% in the eastern Pacific) of the initial positions of our samples occurred south of 30°. Extending the model verification to latitudes north of 30° is left to future work.

Figure 8 shows an example of the exceedance probabilities for the 64-kt wind speed threshold, cumulative over the 5-day forecast period, for Hurricane Irma. The analyzed maximum extent of 64-kt winds, indicated by the thick black contours, are reasonably within the bounds of the probabilistic forecast, except for a mild along-track error. To obtain robust evaluation statistics, we evaluate all the forecasts by analyzing the associated reliability diagrams, multiplicative bias curves, and the maximum threat scores (Wilks 2011).

Figure 9 displays a reliability diagram for the cumulative exceedance wind probabilities using different wind speed thresholds, separated by basin and cumulative time. The results indicate that in general, the exceedance probabilities generated by the FHLO-Super model are reliable at the examined cumulative times and in both basins. We also observe that for probabilities between 0% and 50%, the 34-kt winds are more overdispersive than those for the 50-kt winds, which in turn are more overdispersive than the 64-kt winds. This is likely due to two factors. First, negative biases at $r/r_m \approx 4$ in the axisymmetric wind model cannot be completely eliminated with the shape parameter, though the bias was heavily reduced (not shown). Second, while the TC wind profile is typically dominated by axisymmetric processes near the core, asymmetries often dominate farther away from the core. These asymmetries cannot always be represented by the simple asymmetric model in the aforementioned text. The combined effect of both of these factors is likely to underestimate the radial extent of winds at a particular threshold in each quadrant, where the magnitude of this bias decreases as one moves toward the inner core (i.e., we underestimate r_{34}, and less so for r_{50}, and even less for r_{64}). This observed bias could be addressed by adding more degrees of freedom to the wind model, though this would reduce the simplicity. Using another wind model with more degrees of freedom, such as the radii-CLIPER wind model of Knaff et al. (2007), may also help to alleviate these issues.

The multiplicative bias is defined by $B = \sum_i F_i / \sum_i O_i$, where B is the multiplicative bias, F_i are the forecasted probabilities, and O_i is the observation (0 or 1). The bias B is a measure of whether the average forecast has probabilities that are too large ($B > 1$), or too small ($B < 1$). The average probabilities of FHLO-Super, shown by Fig. 10, are generally too small for the 34- and 50-kt thresholds, and reasonably unbiased for the 64-kt threshold. The underlying reason for $B < 1$ at the 34- and 50-kt thresholds is likely the same reason for the bias in the
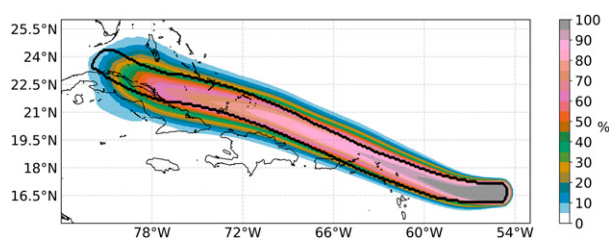

FIG. 8. Probability of wind exceeding 64 kt over a 5-day period, for Hurricane Irma in the Atlantic basin, using 1000 ensemble members bootstrapped from the ECMWF ensemble. Forecast initialized at 0000 UTC 5 Sep 2017. The black contour depicts extent of the observed 64-kt winds.

reliability diagrams; namely, that r_{34} and r_{50} are being underestimated in each quadrant. Regardless, the multiplicative biases at a fixed threshold remain relatively constant in time. This suggests that reducing the bias in the wind model could lead to $B \approx 1$ at all thresholds.

Unfortunately, a large number of grid points across a basin have zero probability and will evaluate to correct nulls, which can strongly influence the bias scores. To remedy this, we use the threat score metric to further evaluate the probabilistic wind forecasts. Correct nulls are not used in the threat score. To calculate the threat score, one defines a threshold probability to determine a categorical forecast (yes or no) and divides the total number of correct forecasts by the sum of the total number of correct forecasts, false positives, and misses. The threat score has a score from 0 (worst) to 1 (best). As a baseline, we perform a homogeneous comparison with NHC's MCP model. NHC only archives the probabilities accumulated over all storms during a particular cycle, which is an issue when combined with the fact that FHLO forecasts only run as far out as the ensemble tracks. Thus, for any forecast, we must discard any samples beyond the lead time at which the FHLO forecast for any existing TC dissipates. This reduces the number of samples we can use to evaluate FHLO. In the Atlantic (eastern Pacific), we used 348 (161) cycles at a lead time of 6 h, to 61 (20) cycles at a lead time of 120 h, using 0000 and 1200 UTC forecasts from 2015 to 2018. We also create a superensemble between FHLO-Super and the MCP model, simply by averaging the wind probabilities of each model, and denote this as FHLO-MCP.

Figure 11 compares the maximum threat scores of FHLO-Super, the MCP model, and FHLO-MCP. The threat scores generally maximize at a threshold probability of ≈30% (not shown). In the Atlantic basin, the MCP model is superior at shorter lead times (≈0–2 days), especially at the 34- and 50-kt thresholds. One reason this may be is that analyses of the initial position and intensity can be explicitly taken into account in the official forecast.
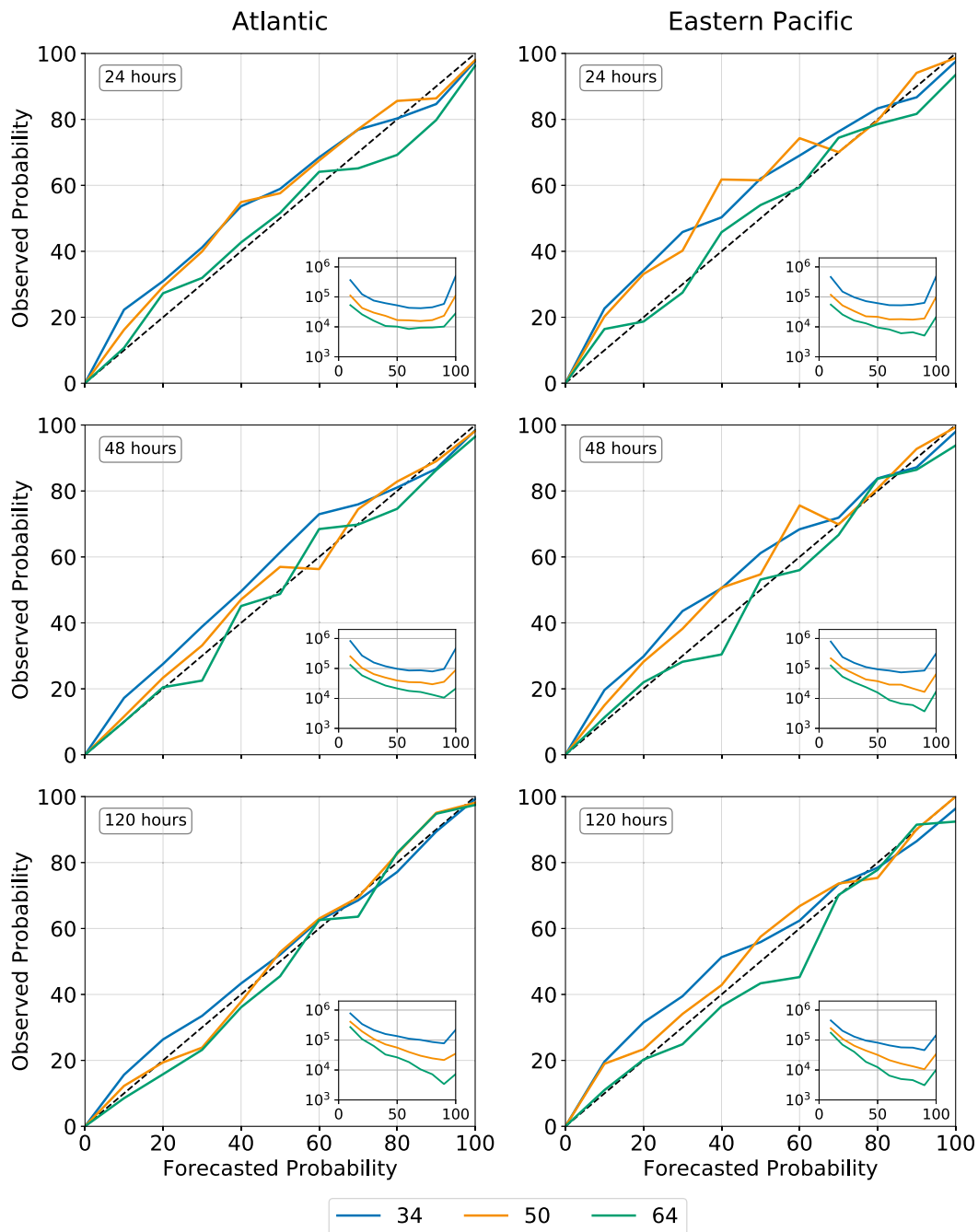
FIG. 9. Reliability diagram for exceedance wind probabilities, cumulative over (top) 24, (middle) 48, and (bottom) 120 h, for the 34-, 50-, and 64-kt thresholds, using the probabilities derived from FHLO-Super. Basins are separated into (left) Atlantic and (right) eastern Pacific. Sample set includes only forecasts where the initial position is south of 30°N. Dashed black line indicates perfect reliability. Insets show the sample size, for each threshold, as a function of forecast probability (days).

On the other hand, FHLO-Super has higher threat scores at longer lead times (≈3 days), regardless of the threshold. In the eastern Pacific basin, FHLO-Super has larger threat scores at nearly all lead times and thresholds. These results suggest that incorporating state-dependent uncertainties could lead to improvements in long-range, probabilistic, pointwise wind probability forecasts. Taking into account the flow-dependent uncertainty will be extremely important if forecast lead times are extended beyond 5 days. Finally, in the Atlantic basin, the FHLO-MCP
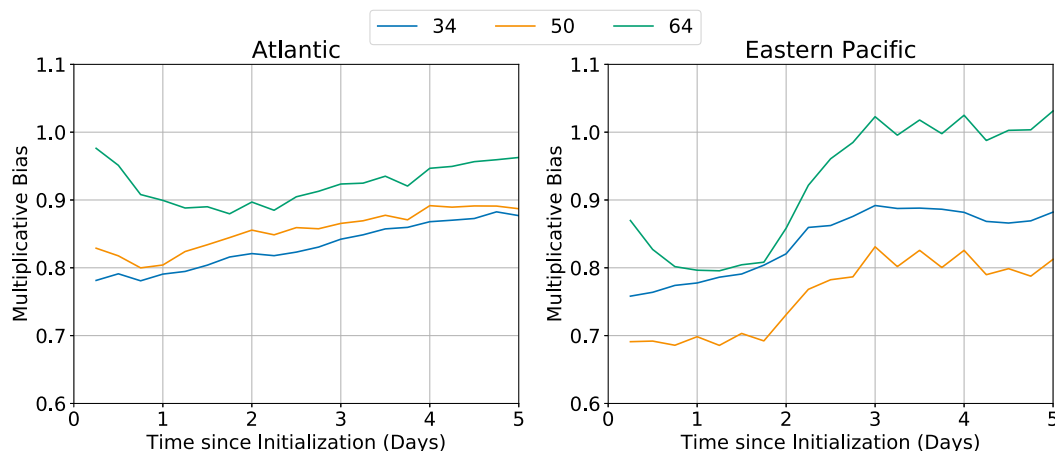
FIG. 10. Multiplicative bias for the FHLO-Super cumulative exceedance wind probabilities, separated into the 34-, 50-, and 64-kt thresholds, and as a function of the time since initialization. Basins are separated into (left) Atlantic and (right) eastern Pacific. Sample set includes only forecasts where the initial position is south of 30°N.

superensemble has better threat scores compared to both FHLO-Super and the MCP model, at all lead times and thresholds. While the sample size of the homogeneous comparison between FHLO-Super and the MCP model may be a bit limited, this provides more evidence that combining multiple models can lead to superior forecasts.

As mentioned previously, ensemble models have the tendency to possess negative biases in the intensity of a TC. This may lead to premature dissipation of a TC, resulting in smaller threat scores from forecast misses. We expect that as the ability to resolve TCs in ensemble

models increases, the threat scores associated with the FHLO-based ensembles will also increase. Regardless, these results suggest that there is much value in incorporating state-dependent uncertainty in TC forecasts.

5. Conclusions

In this study, we developed FHLO, a probabilistic, large-ensemble [$O(1000)$ members], TC prediction framework that estimates forecast uncertainty by leveraging the internal variability of the large-scale environment simulated by a global NWP ensemble. We described a
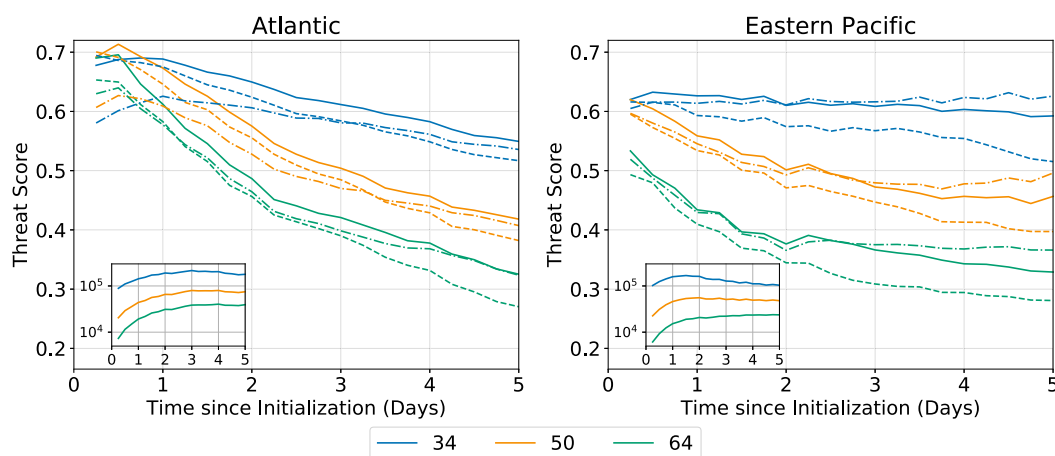

FIG. 11. The maximum threat scores for the FHLO-Super (dot–dashed), MCP (dashed), and FHLO-MCP (solid) cumulative exceedance wind probabilities, separated into the 34-, 50-, and 64-kt thresholds, and as a function of the time since initialization. Basins are separated into (left) Atlantic and (right) eastern Pacific. Sample set includes only forecasts where the initial position is south of 30°N. Insets show the threat score denominator (hits + misses + false alarms) for each threshold, as a function of time since initialization (days), for samples used to evaluate FHLO-MCP.

method to generate synthetic tracks that are statistically similar to those of an ensemble NWP model. We evaluated the intensity forecasts of the simulated TC along each track using the FAST intensity model, and we then used a physically based wind model to estimate the full wind field. We evaluated the model using four years (2015–18) of reforecasts in the Atlantic and eastern Pacific basins.

We show that the FHLO-based large-ensemble intensity forecasts perform comparably with HWRF, an advanced NWP model. We also evaluate the probability of exceedance for the 34-, 50-, and 64-kt wind speed thresholds using reliability curves, multiplicative bias curves, and threat scores. The results suggest that, notwithstanding some slight biases at low wind speed thresholds, the FHLO-based wind forecasts are skillful and reliable. Pointwise wind forecasts using the FHLO framework are particularly skillful at lead times longer than 3 days, and a combination of FHLO-Super and NHC's MCP model was shown to have the highest threat scores across all lead times in the Atlantic. These results suggest that it will be important to better characterize the state-dependent uncertainty to continue to improve our long-range forecast skill.

Probabilistic wind speed forecasts combine the uncertainty in both the track and intensity and provide uncertainty quantification of wind speeds at fixed points. As compared to traditional forecast quantities such as the exact storm center or the maximum wind speed anywhere in the storm, pointwise wind speeds are the more relevant TC hazard metric for assessing wind impacts. The uncertainty of wind speeds at a point could also be translated to location-specific vulnerabilities, as is the goal of the Hurricane Risk Calculator (Vigh et al. 2020). Thus, while the traditional track-intensity dichotomy is useful for evaluating and improving our models, it is not the most useful and pertinent metric to the public.

It is also important to stress that large ensembles [$O(1000)$ ensemble members] are necessary to provide accurate forecasts of pointwise wind speed probability distributions. This is one of the key advantages of the large-ensemble framework. Furthermore, unlike the model described in DeMaria et al. (2009), the FHLO framework quantifies forecast uncertainty by incorporating the internal variability in the large-scale environment. This is achieved by including uncertainty in the TC's track, uncertainty in the dynamic and thermodynamic environments, and uncertainty in the initial conditions. The synthetic tracks of the model are generated by sampling from the dispersion among TC tracks of an ensemble NWP model. The dynamic and thermodynamic fields of the ensemble NWP system are also used to generate realistic perturbations to TC-relevant environmental quantities, such as the saturation entropy deficit and vertical wind shear. Initial condition uncertainty is accounted for by incorporating stochastic perturbations to the initial intensity and inner-core moisture.

There is also uncertainty in the observations that are used to evaluate the model. While we evaluate the wind speed probabilities by assuming that all points within the estimated maximum radial extent of a particular wind speed threshold, this may likely not be the case. A better method to evaluate wind speed probabilities is to only use wind observations over land, which are arguably more objective than best track estimates of intensity. Undertaking such an evaluation could be an area of future work.

It is worth discussing the various sources of error and limitations of the FHLO framework. First, while we provide evidence that FAST can reasonably model and forecast TCs, the FAST equations are still an imperfect and idealized representation of the evolution of the intensity of a TC. We expect the intensity module to work best for axisymmetric, surface-flux-driven, mature TCs, and thus less so for tropical disturbances or cyclones that are undergoing extratropical transition. A statistical bias correction could be applied to the intensity component for weak disturbances in light of this issue. Furthermore, in the FAST framework, we consider midlevel ventilation as the only process that dries out the inner core of a TC. The eddy-entropy flux into the eyewall is then approximated to vary linearly with the product of the environmental saturation entropy deficit and the vertical wind shear (Tang and Emanuel 2012). Any departures from this approximation will affect the intensity model. Uncertainty and error can also arise from the methods developed to calculate environmental quantities. The definition of vertical wind shear that is relevant to the TC is problematic, since errors can be introduced through $r*$, as well as the vertical levels by which to calculate shear. In this study, we fixed $r* = 400$ km and used the 250–850-hPa vertical wind shear. These are perhaps not the best approximations, and it may be the case that the "optimal" $r*$ and vertical levels by which to calculate vertical wind shear can vary from storm to storm.

Since TCs are primarily driven by thermodynamic disequilibrium between the sea surface and boundary layer air (Emanuel 1986), accurate observations of the ocean are important to properly calculate environmental quantities, such as the potential intensity. For this study, errors in the estimation of these quantities could have been introduced with the smoothing operators, which were necessary to remove the TC's influence on the thermodynamic and dynamic fields. Furthermore, we parameterize ocean mixing from upwelling by using climatological mixed layer depth and sub–mixed layer

thermal stratification, which effectively smooths out any high-frequency variability in the ocean. This could lead to large errors in intensity forecasts (Emanuel et al. 2004). Future work will incorporate information from a real-time ocean model.

FHLO can also be used to forecast the probability of rapid intensification (RI). While a full evaluation of the skillfulness of the FHLO-based ensemble forecasts to predict RI is left for future work, preliminary evaluation of probabilistic rapid intensification forecasts, regardless of the intensity of the storm, shows some promise (not shown). The evaluation results of the track and intensity forecasts also suggest that combining many more ensemble models, such as the Met Office and Japan Meteorological Agency ensemble systems, could lead to more skillful forecasts, as suggested by Yamaguchi et al. (2012) and Titley et al. (2020). In addition, while we only used one model to predict TC intensity in this study, using a variety of intensity models to generate intensity forecasts may improve the overall model. The only restriction is that the intensity model must be computationally inexpensive. This principle applies equally to the wind field model.

Finally, it is important to note that the skill of the FHLO-based ensemble is derived from the accuracy of the ensemble NWP system. While it is clear that errors in representing the environmental fields will significantly affect the intensity component, errors in representing the TC in the NWP model itself can also affect the FHLO-based ensemble. This can happen when a TC dissipates too early in the ensemble prediction model, such that the track model will generate tracks that also dissipate too early. Though the skill of FHLO is significantly coupled to the quality of the ensemble system, it is likely that the skill of FHLO will progress along with advancements in ensemble NWP models. This is arguably the most intriguing aspect of FHLO: it can be viewed as a framework for bootstrapping an ensemble NWP model.

APPENDIX

Model Setup Details

a. Inversion on a low-resolution grid

In this study, we remove winds associated with the TC by performing ''vortex surgery'' on both the ERA5 analysis fields and the forecast ensemble fields, as in Davis et al. (2008). While the analysis fields can be obtained at high grid resolution, the grid resolution (0.5°) of ensemble numerical weather prediction models is inadequate to properly resolve the vorticity maxima of a TC. One way around this is to interpolate between grid points. However, interpolation schemes must always assume some structure about the field (i.e., linear, cubic, etc.). Using spherical harmonics, we can extend the number of basis functions to an arbitrary higher order by assuming that the power spectrum obeys a well-defined power scaling law. Defining L as the truncation order of the original field, L' as the truncation order of the higher resolution field, $S(l')$ as the desired power at degree l', and n as a random coefficient with unit variance, then, the superresoluted field f is

$$f(\theta,\phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} f_l^m Y_l^m(\theta,\phi)$$
$$+ \sum_{l'=l+1}^{L'} \sum_{m'=-l'}^{l'} \left[\sqrt{S(l')}\, n \right] Y_{l'}^{m'}(\theta,\phi),$$

where S is determined from the power spectrum at the extended spherical harmonics. For simplicity, we use an exponentially decaying fit to the power spectrum and use the fit to extend to higher order of harmonics. Note that this smooth interpolation can also be performed without assuming any structure about the power spectrum, as long as the original resolution is high enough such that the amplitudes of the higher-order spherical harmonics are sufficiently small. In fact, we can even set the power of the higher-order spherical harmonics to zero with relatively little consequence on the resulting environmental wind fields.

b. Perturbations to observed intensity

To generate perturbations to the observed intensity, we model the intensity, $V(t)$, as a Gaussian process with the time-varying observations as the mean and a constant covariance kernel:

$$C_V(t_1, t_2) = \sigma_V^2 \exp(-|t_1 - t_2|/T),$$

where t_1 and t_2 are discrete time points in the observed history, $T = 1$ day, and σ^2 is the variance that represents the uncertainty associated with the observations. We write the Karhunen–Loeve expansion of the intensity as $V(t) = \mu_V(t) + \sum_{i=1}^{n} \sqrt{\lambda_i} \psi_i \theta_i$, where n is the total stochastic dimension, θ_i is a standard normal random variable, and λ_i and ψ_i are the ordered eigenvalues and eigenvectors with respect to the covariance kernel C_V. We only consider fluctuations of the intensity on time scales of the observations (every 6 h), such that we take $n = 10$ to remove any high-frequency variability. Finally, we take σ_v^2 as a piecewise constant function:

$$\sigma^2(v) = \begin{cases} 5, & \text{if} \quad v < 32 \text{ m s}^{-1} \\ 10, & \text{if} \quad v \geq 33 \text{ m s}^{-1} \end{cases},$$

where v is the observed intensity. These values of $\sigma^2(v)$ were chosen as simple approximations to the uncertainty distributions shown in Landsea and Franklin (2013).

REFERENCES

Bougeault, P., and Coauthors, 2010: The THORPEX Interactive Grand Global Ensemble. *Bull. Amer. Meteor. Soc.*, **91**, 1059–1072, https://doi.org/10.1175/2010BAMS2853.1.

Bryan, G. H., 2008: On the computation of pseudoadiabatic entropy and equivalent potential temperature. *Mon. Wea. Rev.*, **136**, 5239–5245, https://doi.org/10.1175/2008MWR2593.1.

Chavas, D. R., and K. Emanuel, 2010: A QuikSCAT climatology of tropical cyclone size. *Geophys. Res. Lett.*, **37**, L18816, https://doi.org/10.1029/2010GL044558.

——, and N. Lin, 2016: A model for the complete radial structure of the tropical cyclone wind field. Part II: Wind field variability. *J. Atmos. Sci.*, **73**, 3093–3113, https://doi.org/10.1175/JAS-D-15-0185.1.

——, ——, and K. Emanuel, 2015: A model for the complete radial structure of the tropical cyclone wind field. Part I: Comparison with observed structure. *J. Atmos. Sci.*, **72**, 3647–3662, https://doi.org/10.1175/JAS-D-15-0014.1.

Copernicus Climate Change Service, 2017: ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store, accessed 16 July 2019, https://cds.climate.copernicus.eu/cdsapp#!/home.

Davis, C., C. Snyder, and A. C. Didlake Jr., 2008: A vortex-based perspective of eastern Pacific tropical cyclone formation. *Mon. Wea. Rev.*, **136**, 2461–2477, https://doi.org/10.1175/2007MWR2317.1.

DeMaria, M., and J. Kaplan, 1994: A Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting*, **9**, 209–220, https://doi.org/10.1175/1520-0434(1994)009<0209:ASHIPS>2.0.CO;2.

——, J. A. Knaff, R. Knabb, C. Lauer, C. R. Sampson, and R. T. DeMaria, 2009: A new method for estimating tropical cyclone wind speed probabilities. *Wea. Forecasting*, **24**, 1573–1591, https://doi.org/10.1175/2009WAF2222286.1.

——, and Coauthors, 2013: Improvements to the operational tropical cyclone wind speed probability model. *Wea. Forecasting*, **28**, 586–602, https://doi.org/10.1175/WAF-D-12-00116.1.

——, C. R. Sampson, J. A. Knaff, and K. D. Musgrave, 2014: Is tropical cyclone intensity guidance improving? *Bull. Amer. Meteor. Soc.*, **95**, 387–398, https://doi.org/10.1175/BAMS-D-12-00240.1.

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

——, 2000: A statistical analysis of tropical cyclone intensity. *Mon. Wea. Rev.*, **128**, 1139–1152, https://doi.org/10.1175/1520-0493(2000)128<1139:ASAOTC>2.0.CO;2.

——, 2011: Global warming effects on U.S. hurricane damage. *Wea. Climate Soc.*, **3**, 261–268, https://doi.org/10.1175/WCAS-D-11-00007.1.

——, 2013: Downscaling CMIP5 climate models shows increased tropical cyclone activity over the 21st century. *Proc. Natl. Acad. Sci. USA*, **110**, 12 219–12 224, https://doi.org/10.1073/pnas.1301293110.

——, 2017: A fast intensity simulator for tropical cyclone risk analysis. *Nat. Hazards*, **88**, 779–796, https://doi.org/10.1007/s11069-017-2890-7.

——, and F. Zhang, 2016: On the predictability and error sources of tropical cyclone intensity forecasts. *J. Atmos. Sci.*, **73**, 3739–3747, https://doi.org/10.1175/JAS-D-16-0100.1.

——, and ——, 2017: The role of inner-core moisture in tropical cyclone predictability and practical forecast skill. *J. Atmos. Sci.*, **74**, 2315–2324, https://doi.org/10.1175/JAS-D-17-0008.1.

——, C. DesAutels, C. Holloway, and R. Korty, 2004: Environmental control of tropical cyclone intensity. *J. Atmos. Sci.*, **61**, 843–858, https://doi.org/10.1175/1520-0469(2004)061<0843:ECOTCI>2.0.CO;2.

——, R. Sundararajan, and J. Williams, 2008: Hurricanes and global warming: Results from downscaling IPCC AR4 simulations. *Bull. Amer. Meteor. Soc.*, **89**, 347–368, https://doi.org/10.1175/BAMS-89-3-347.

Galarneau, T. J., and C. A. Davis, 2013: Diagnosing forecast errors in tropical cyclone motion. *Mon. Wea. Rev.*, **141**, 405–430, https://doi.org/10.1175/MWR-D-12-00071.1.

Gentry, M. S., and G. M. Lackmann, 2010: Sensitivity of simulated tropical cyclone structure and intensity to horizontal resolution. *Mon. Wea. Rev.*, **138**, 688–704, https://doi.org/10.1175/2009MWR2976.1.

Gneiting, T., and A. E. Raftery, 2005: Weather forecasting with ensemble methods. *Science*, **310**, 248–249, https://doi.org/10.1126/science.1115255.

Goerss, J. S., 2000: Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Wea. Rev.*, **128**, 1187–1193, https://doi.org/10.1175/1520-0493(2000)128<1187:TCTFUA>2.0.CO;2.

——, 2007: Prediction of consensus tropical cyclone track forecast error. *Mon. Wea. Rev.*, **135**, 1985–1993, https://doi.org/10.1175/MWR3390.1.

Gopalakrishnan, S. G., F. Marks Jr., X. Zhang, J.-W. Bao, K.-S. Yeh, and R. Atlas, 2011: The experimental HWRF system: A study on the influence of horizontal resolution on the structure and intensity changes in tropical cyclones using an idealized framework. *Mon. Wea. Rev.*, **139**, 1762–1784, https://doi.org/10.1175/2010MWR3535.1.

Hamill, T. M., J. S. Whitaker, M. Fiorino, and S. G. Benjamin, 2011: Global ensemble predictions of 2009's tropical cyclones initialized with an ensemble Kalman filter. *Mon. Wea. Rev.*, **139**, 668–688, https://doi.org/10.1175/2010MWR3456.1.

Hersbach, H., 2016: The ERA5 atmospheric reanalysis. *2016 Fall Meeting*, San Francisco, CA, Amer. Geophys. Union, Abstract NG33D-01.

Judt, F., S. S. Chen, and J. Berner, 2016: Predictability of tropical cyclone intensity: Scale-dependent forecast error growth in high-resolution stochastic kinetic-energy backscatter ensembles. *Quart. J. Roy. Meteor. Soc.*, **142**, 43–57, https://doi.org/10.1002/qj.2626.

Kieu, C. Q., and Z. Moon, 2016: Hurricane intensity predictability. *Bull. Amer. Meteor. Soc.*, **97**, 1847–1857, https://doi.org/10.1175/BAMS-D-15-00168.1.

Knaff, J. A., C. R. Sampson, M. DeMaria, T. P. Marchok, J. M. Gross, and C. J. McAdie, 2007: Statistical tropical cyclone wind radii prediction using climatology and persistence. *Wea. Forecasting*, **22**, 781–791, https://doi.org/10.1175/WAF1026.1.

——, ——, and G. Chirokova, 2017: A global statistical–dynamical tropical cyclone wind radii forecast scheme. *Wea. Forecasting*, **32**, 629–644, https://doi.org/10.1175/WAF-D-16-0168.1.

Landsea, C. W., and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592, https://doi.org/10.1175/MWR-D-12-00254.1.

Levitus, S., 1982: *Climatological Atlas of the World Ocean*. NOAA Prof. Paper 13, 173 pp. and 17 microfiche.

Liu, Q., T. Marchok, H.-L. Pan, M. Bender, and S. Lord, 2000: Improvements in hurricane initialization and forecasting at NCEP with global and regional (GFDL) models. NWS Tech. Procedures Bull. 472, 7 pp.

Majumdar, S. J., and P. M. Finocchio, 2010: On the ability of global ensemble prediction systems to predict tropical cyclone track probabilities. *Wea. Forecasting*, **25**, 659–680, https://doi.org/10.1175/2009WAF2222327.1.

Marks, D. G., 1992: The beta and advection model for hurricane track forecasting. NWS Tech. Memo. NWS NMC 70, 89 pp., https://repository.library.noaa.gov/view/noaa/7184.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, https://doi.org/10.1002/qj.49712252905.

National Centers for Environmental Prediction, 2015: NCEP GFS 0.25 degree global forecast grids historical archive. National Center for Atmospheric Research, Computational and Information Systems Laboratory, accessed 16 August 2019, https://doi.org/10.5065/D65D8PWK.

——, 2008: THORPEX Interactive Grand Global Ensemble (TIGGE) model tropical cyclone track data. National Center for Atmospheric Research, Computational and Information Systems Laboratory, accessed 17 May 2019, https://doi.org/10.5065/D6GH9GSZ.

Pielke, R. A., J. Gratz, C. W. Landsea, D. Collins, M. A. Saunders, and R. Musulin, 2008: Normalized hurricane damage in the United States: 1900–2005. *Nat. Hazards Rev.*, **9**, 29–42, https://doi.org/10.1061/(ASCE)1527-6988(2008)9:1(29).

Rappaport, E. N., and Coauthors, 2009: Advances and challenges at the National Hurricane Center. *Wea. Forecasting*, **24**, 395–419, https://doi.org/10.1175/2008WAF2222128.1.

Reynolds, R. W., V. F. Banzon, and N. C. Program, 2008: NOAA Optimum Interpolation 1/4 Degree Daily Sea Surface Temperature (OISST) analysis, version 2. NOAA National Centers for Environmental Information, accessed 13 August 2019, https://doi.org/10.7289/V5SQ8XB5.

Sampson, C. R., and A. J. Schrader, 2000: The automated tropical cyclone forecasting system (version 3.2). *Bull. Amer. Meteor. Soc.*, **81**, 1231–1240, https://doi.org/10.1175/1520-0477(2000)081<1231:TATCFS>2.3.CO;2.

Simpson, R., and R. Riehl, 1958: Mid-tropospheric ventilation as a constraint on hurricane development and maintenance. Preprints, *Tech. Conf. on Hurricanes*, Miami Beach, FL, Amer. Meteor. Soc., D4-1–D4-10.

Sippel, J. A., and F. Zhang, 2008: A probabilistic analysis of the dynamics and predictability of tropical cyclogenesis. *J. Atmos. Sci.*, **65**, 3440–3459, https://doi.org/10.1175/2008JAS2597.1.

Swinbank, R., and Coauthors, 2016: The TIGGE project and its achievements. *Bull. Amer. Meteor. Soc.*, **97**, 49–67, https://doi.org/10.1175/BAMS-D-13-00191.1.

Tang, B., and K. Emanuel, 2010: Midlevel ventilation's constraint on tropical cyclone intensity. *J. Atmos. Sci.*, **67**, 1817–1830, https://doi.org/10.1175/2010JAS3318.1.

——, and ——, 2012: A ventilation index for tropical cyclones. *Bull. Amer. Meteor. Soc.*, **93**, 1901–1912, https://doi.org/10.1175/BAMS-D-11-00165.1.

Titley, H. A., R. L. Bowyer, and H. L. Cloke, 2020: A global evaluation of multi-model ensemble tropical cyclone track probability forecasts. *Quart. J. Roy. Meteor. Soc.*, **146**, 531–545, https://doi.org/10.1002/qj.3712.

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting*, **8**, 379–398, https://doi.org/10.1175/1520-0434(1993)008<0379:OEPATN>2.0.CO;2.

Van Sang, N., R. K. Smith, and M. T. Montgomery, 2008: Tropical-cyclone intensification and predictability in three dimensions. *Quart. J. Roy. Meteor. Soc.*, **134**, 563–582, https://doi.org/10.1002/qj.235.

Vigh, J. L., and Coauthors, 2020: The Hurricane Risk Calculator: Working toward enhancing our nation's readiness, responsiveness, and resilience to hurricanes through probabilistic risk frameworks for evacuation decision support. *Eighth Symp. on Building a Weather-Ready Nation: Enhancing Our Nation's Readiness, Responsiveness, and Resilience to High Impact Weather Events*, Boston, MA, Amer. Meteor. Soc., 5.5, https://ams.confex.com/ams/2020Annual/webprogram/Paper370408.html.

Vijaya Kumar, T., T. Krishnamurti, M. Fiorino, and M. Nagata, 2003: Multimodel superensemble forecasting of tropical cyclones in the Pacific. *Mon. Wea. Rev.*, **131**, 574–583, https://doi.org/10.1175/1520-0493(2003)131<0574:MSFOTC>2.0.CO;2.

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

Williford, C. E., T. N. Krishnamurti, R. C. Torres, S. Cocke, Z. Christidis, and T. S. Vijaya Kumar, 2003: Real-time multimodel superensemble forecasts of Atlantic tropical systems of 1999. *Mon. Wea. Rev.*, **131**, 1878–1894, https://doi.org/10.1175/2571.1.

Yamaguchi, M., T. Nakazawa, and S. Hoshino, 2012: On the relative benefits of a multi-centre grand ensemble for tropical cyclone track prediction in the western North Pacific. *Quart. J. Roy. Meteor. Soc.*, **138**, 2019–2029, https://doi.org/10.1002/qj.1937.

Zhang, F., and J. A. Sippel, 2009: Effects of moist convection on hurricane predictability. *J. Atmos. Sci.*, **66**, 1944–1961, https://doi.org/10.1175/2009JAS2824.1.

——, and D. Tao, 2013: Effects of vertical wind shear on the predictability of tropical cyclones. *J. Atmos. Sci.*, **70**, 975–983, https://doi.org/10.1175/JAS-D-12-0133.1.