Naval Research Laboratory
Washington, DC 20375-5320

NRL/7530/FR--2022/1

Validation Test Report: Coupled Ocean/Atmosphere Mesoscale Prediction System—Tropical Cyclone Ensemble (COAMPS-TC Ensemble) v2021

WILLIAM KOMAROMI
JON MOSKAITIS
ALEX REINECKE

Atmospheric Dynamics & Prediction Branch
Marine Meteorology Division

JAMES DOYLE

Senior Scientist for Mesoscale Meteorology
Marine Meteorology Division

CHARLES SKUPNIEWICZ
ROGER STOCKE
CAREY DICKERMAN

Fleet Numerical Meteorology and Oceanography Center
Monterey, CA

January 27, 2022

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
1/27/2022	NRL Formal Report	10/1/2020	11/8/2021

4. TITLE AND SUBTITLE

Validation Test Report: Coupled Ocean/Atmosphere Mesoscale Prediction System—Tropical Cyclone Ensemble (COAMPS-TC Ensemble) v2021

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
		100001802004 0010
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
0603207N		

6. AUTHOR(S)

William Komaromi, James Doyle, Jon Moskaitis, Alex Reinecke, Charles Skupniewicz,* Roger Stocke,* and Carey Dickerman*

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Naval Research Laboratory Monterey 7 Grace Hopper Ave, Monterey CA 93943-5598 Fleet Numerical Meteorology & Oceanography Center 7 Grace Hopper Ave, Monterey CA 93943-5598	NRL/7530/FR--2022/1

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Ocean, Atmosphere, and Space Research Division Office of Naval Research, Code 322MM 875 North Randolph Street, Suite 1425 Arlington, VA 22203-1995	ONR	

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

*Fleet Numerical Meteorology and Oceanography Center, 7 Grace Hopper Ave, Monterey, VA 93943-5598

14. ABSTRACT

In this transition, the 11-member COAMPS-TC ensemble is upgraded from v2018 to v2021. As detailed in this report, this transition constitutes a significant upgrade. Included with this transition are: adjusted synoptic-scale initial time and lateral boundary perturbation magnitudes; updates to the shallow cumulus parameterization on grids 1 and 2; modifications to *tcinit* to produce smaller, more realistic tropical cyclones (TCs); the implementation of graupel-radiation interaction; initialization off of 0.25 deg GFS (versus 0.50 deg GFS as is presently done); GFS downscaling for weak TCs to produce a more realistic, less symmetric vortex for weaker storms; an updated surface drag coefficient; an improved 1-dimensional sea surface temperature cooling parameterization; changes to the interaction between grids 2 and 3 with the grid 1 blendzone; and increased diffusion in the first 6 h of the forecast for weak TCs. Versions v2021 and v2018 are compared against one another using a sample of 412 retrospective forecast cases for the unperturbed control member and 180 cases run using the full 11-member ensemble. Model upgrades in this transition collectively decrease mean absolute error (MAE) and bias for both track and intensity, improve rapid intensification (RI) statistics, improve the pressure-wind relationship and the intensity forecast distribution, and improve the 34, 50, and 64 kt wind radii. Probabilistic verification metrics show, for example, that uncertainty discrimination for track is slightly degraded but uncertainty discrimination for intensity is significantly improved.

The ensemble has also been modified such that it can now run Invests, which are often high on JTWC's priority list. Finally, the graphics suite has also been updated to use Python 3 and Cartopy (v2018 used Python 2 and Basemap), which produces more user-friendly graphical products.

15. SUBJECT TERMS

COAMPS-TC tropical cyclones numerical weather prediction ensemble forecasting meteorology

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U/U	U/U	U/U	U/U	29

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
William A. Komaromi	(831) 656-4038

This page intentionally left blank

CONTENTS

1. INTRODUCTION 1
2. MODEL DESCRIPTION AND UPGRADES IN v2021 1
   2.1 Model Overview 1
   2.2 Upgrades to Model Physics 2
   2.3 Improved Tropical Cyclone Initialization 3
   2.4 Ensemble Configuration Upgrades 3
3. ENSEMBLE CONTROL MEMBER VERIFICATION 6
   3.1 Overview of Cases 6
   3.2 Results for all Basins 7
   3.3 West Pacific results 12
4. FULL 11-MEMBER ENSEMBLE FORECASTS 15
   4.1 Overview of Cases 15
   4.1 Deterministic Validation of the Ensemble Mean 16
   4.3 Probabilistic Verification 18
5. SUMMARY AND CONCLUSIONS 23
6. REFERENCES 24

This page intentionally left blank

EXECUTIVE SUMMARY

In this transition, the 11-member COAMPS-TC ensemble is upgraded from v2018 to v2021. As detailed in this report, this transition constitutes a significant upgrade. Included with this transition are: adjusted synoptic-scale initial time and lateral boundary perturbation magnitudes; updates to the shallow cumulus parameterization on grids 1 and 2; modifications to *tcinit* to produce smaller, more realistic tropical cyclones (TCs); the implementation of graupel-radiation interaction; initialization off of 0.25 deg GFS (versus 0.50 deg GFS as is presently done); GFS downscaling for weak TCs to produce a more realistic, less symmetric vortex for weaker storms; an updated surface drag coefficient; an improved 1-dimensional sea surface temperature cooling parameterization; changes to the interaction between grids 2 and 3 with the grid 1 blendzone; and increased diffusion in the first 6 h of the forecast for weak TCs. Versions v2021 and v2018 are compared against one another using a sample of 412 retrospective forecast cases for the unperturbed control member and 180 cases run using the full 11-member ensemble. Model upgrades in this transition collectively decrease mean absolute error (MAE) and bias for both track and intensity, improve rapid intensification (RI) statistics, improve the pressure-wind relationship and the intensity forecast distribution, and improve the 34, 50, and 64 kt wind radii. Probabilistic verification metrics show, for example, that uncertainty discrimination for track is slightly degraded but uncertainty discrimination for intensity is significantly improved.

The ensemble has also been modified such that it can now run Invests, which are often high on JTWC's priority list. Finally, the graphics suite has also been updated to use Python 3 and Cartopy (v2018 used Python 2 and Basemap), which produces more user-friendly graphical products.

This page intentionally left blank

VALIDATION TEST REPORT: COUPLED OCEAN/ATMOSPHERE MESOSCALE PREDICTION SYSTEM–TROPICAL CYCLONE ENSELMBLE (COAMPS-TC ENSEMBLE) v2021

1. INTRODUCTION

The Coupled Ocean-Atmosphere Mesoscale Prediction System for Tropical Cyclones (COAMPS-TC®) is a dynamical tropical cyclone (TC) mesoscale model with deterministic track and intensity predictions competitive with the best models in the world (Doyle et al. 2020, Masters 2020). The COAMPS-TC deterministic and 11-member ensemble systems are amongst the most heavily leveraged sources of guidance to forecasters at the Joint Typhoon Warning Center (JTWC) for predicting tropical cyclone (TC) position; intensity, including the probability of rapid intensification (RI); wind radii; and structure. The COAMPS-TC ensemble leverages the skill of the deterministic COAMPS-TC core model while accounting for sources of uncertainty in the environmental initial conditions, environmental boundary conditions, the initial vortex, and the model physics to produce a probabilistic forecast. The version 2018 (v2018) of the ensemble was transitioned from NRL to FNMOC in October 2018. This transition to v2021 constitutes a significant upgrade, with a number of improvements to the core COAMPS-TC code now implemented into the ensemble Cylc workflow in addition to a number of ensemble-specific improvements. Details of these changes are outlined in section 2.

2. MODEL DESCRIPTION AND UPGRADES IN v2021

2.1 Model Overview

The COAMPS-TC model features a nonhydrostatic dynamical core and comprises analysis, initialization, and forecast model sub-components (Doyle et al. 2014; 2012). The suite of physical parameterizations for the atmospheric model includes representations for cloud microphysics, boundary layer and free-atmospheric turbulent mixing, surface fluxes, radiation, and deep and shallow convection. The current operational deterministic COAMPS-TC version (v2021) uses a fixed outer grid mesh at 36 km horizontal grid spacing and two storm-following inner grid meshes at 12 km and 4 km grid spacing. There are a total of 7 different outer meshes for the various TC basins around the world. The atmospheric model uses 40 vertical levels with a top near 10 hPa. Initial conditions (ICs) and boundary conditions (BCs) are provided by either the National Oceanic and Atmospheric Administration (NOAA) Global Forecast System (GFS) or the Navy Global Environmental Model (NAVGEM) system.

Manuscript approved January 12, 2022.

The COAMPS-TC ensemble is a probabilistic tropical cyclone forecasting system that has been developed to provide operational, real-time guidance to JTWC (Komaromi et al. 2021). This model accounts for key uncertainties associated with model initial and boundary conditions, and is globally relocatable and configurable in the same way as the deterministic COAMPS-TC model. Details regarding the current configurations of the 2021 COAMPS-TC ensemble appear in Table 1. COAMPS-TC ensemble mean track and intensity forecasts are as accurate as or more accurate than those of the deterministic COAMPS-TC model. The ensemble has been developed so that each of the perturbed members represents an equally likely outcome. A characteristic of a well-configured ensemble is to be able to distinguish between high and low uncertainty forecast scenarios. Forecasts with greater (lesser) uncertainty should be associated with greater (lower) ensemble spread as well as higher (lower) ensemble mean error (ME) when averaged over a sufficiently-large sample. The COAMPS-TC ensemble has been demonstrated to be capable of distinguishing between high and low uncertainty forecast scenarios.

Table 1—Configuration Details for the 2021 Version of the COAMPS-TC Ensemble

Atmospheric Horizontal Resolution	36, 12, 4 km with the 12 and 4 km meshes following the storm
Atmospheric Vertical Resolution	40 vertical levels with the model top at 10 hPa
Grid Mesh Sizes	36 km: 361x191; 12 km: 151x151; 4 km: 226x226
Radiation Parameterization	Fu and Liou (1993)
Microphysics Parameterization	Rutledge and Hobbs (1983) modified by J. Schmidt
Cumulus Parameterization	Kain and Fritsch (1993) on the 36 and 12 km meshes only
Boundary Layer Parameterization	Mellor and Yamada (1982); Bougeault (1985)
Surface Layer Parametrization	Louis (1979); Wang et al. (2002)
Drag for High Winds and Dissipative Heating	Jin et al. (2007), perturbed for members above 25 m s^{-1}
Shallow Convection	Tiedtke (1989)
Ensemble Size	10 perturbed members + 1 unperturbed control member
Large-Scale Perturbations	Static, based upon climatological errors
Vortex-Scale Perturbations (TCs >= 60 kt)	Perturb initial intensity based on analysis errors, constant RH

2.2 Upgrades to Model Physics

A significant number of changes were implemented in the COAMPS-TC ensemble in this upgrade, which include model physics improvements, changes to the initialization, and general model configuration

changes. Note that all of the changes in this section and the following section are already in the v2021 uncoupled deterministic model.

The first model physics improvement included in this upgrade is a replacement of the outdated shallow cumulus parameterization (which was originally ported from NAVGEM) with Tiedtke shallow cumulus on grids 1 and 2. The shallow cumulus scheme transports heat and moisture out of the planetary boundary layer (PBL) into the free troposphere above it. We found that the Tiedtke shallow cumulus scheme produces more realistic cooling and drying in the PBL when compared against GFS and ECMWF analyses, and more realistic moistening and warming in the middle levels.

A second physics upgrade implemented is interaction between the graupel produced by the NRL single-moment microphysics scheme and the Fu-Liou radiation. This change allows the graupel to absorb incoming shortwave radiation and emit longwave radiation, which is particularly relevant to the dynamics of the inner core of the TC. Previously, the radiation scheme was completely unaware of the existence of graupel.

The surface drag coefficient, C_d, was updated from icd12 to icd15, which features a more significant reduction in C_d at higher wind speeds. This allows COAMPS-TC to produce stronger storms than previous capabilities allowed. Without this change, COAMPS-TC could only rarely predict Category 5 cyclones. In conjunction with the surface drag coefficient update, the 1-D column sea surface temperature (SST) cooling parameterization has also been adjusted. In the absence of full 3-D coupling, this parametrization has been found to be quite effective at weakening large, slow moving strong storms.

2.3 Improved Tropical Cyclone Initialization

Several upgrades to the model initialization were also implemented. In order to address a long-standing large bias in terms of the size of the wind field associated with COAMPS-TC cyclones, a bias correction was applied to the initial vortex generated by the TC initialization package “tcinit”. It is found that these changes reduce the large bias meaningfully at earlier lead times and improve the bias through 120 h, albeit with reduced impact at longer lead times. The large bias in the COAMPS-TC initial vortex cannot be totally eliminated due to restrictions on the minimum size of the radius of maximum winds (RMW) at 4 km horizontal resolution and a limit on the radial decay of the wind field outside the RMW. These limits are used to prevent a rapid spin-down of the initial vortex due to the grid discretization.

One of the more significant changes in this upgrade is that tcinit is no longer used to produce a synthetic initial vortex for TCs with intensity less than or equal to 55 kt in the JTWC or NHC warning message. It was found that the synthetic vortex was unrealistically symmetric and vertically aligned for weaker storms; this appeared to be detrimental to the forecast. Instead, the initial vortex state for these weaker storms is simply downscaled from the GFS analysis to the COAMPS-TC grids. The ensemble has also been modified to use 0.25 deg GFS initial conditions and forecast boundary conditions by default, although the option to use 0.50 deg GFS still exists in the event that 0.25 deg GFS is not available. Using the higher resolution GFS analysis has been found to be particularly beneficial for weaker TCs in which tcinit is no longer used, as the 0.25 deg GFS analysis better captures maxima in the wind field, minima in the pressure field, and small-scale asymmetries to the vortex as compared to the 0.50 deg GFS analysis.

In order to address an erroneous spin-up bias that occurs for certain weak storms during the first 6 h of the forecast, increased diffusion has been added for just the first 6 h of the forecast (only for TCs less than or equal to 55 kt intensity at the initial time).

2.4 Ensemble Configuration Upgrades

A number of model configuration changes were implemented in this transition. First, the synoptic-scale initial-time perturbation scale factor (fcp_pert_ic) and lateral boundary condition perturbation scale factor (fcp_pert_bc) magnitudes have been adjusted. fcp_pert_ic was reduced from 1.00 to 0.75, and fcp_pert_bc reduced from 1.75 to 1.00. An example of the effect of this change on the ensemble

track distribution on a high-uncertainty storm, Hurricane Dorian (2019); and on a low-uncertainty storm, Typhoon Faxai (2019), appears in Fig. 1. For both cases, there is a moderate reduction in track spread at all lead times. This change was made in response to several cases identified in which the ensemble produced seemingly unrealistic results, such as the TC tracking directly into the subtropical ridge. This problem appears to be mitigated with the new reduced perturbation scale parameters. Note that the static ensemble perturbation matrices that the ensemble randomly draws from are unchanged. This change only affects the perturbation scale factors.

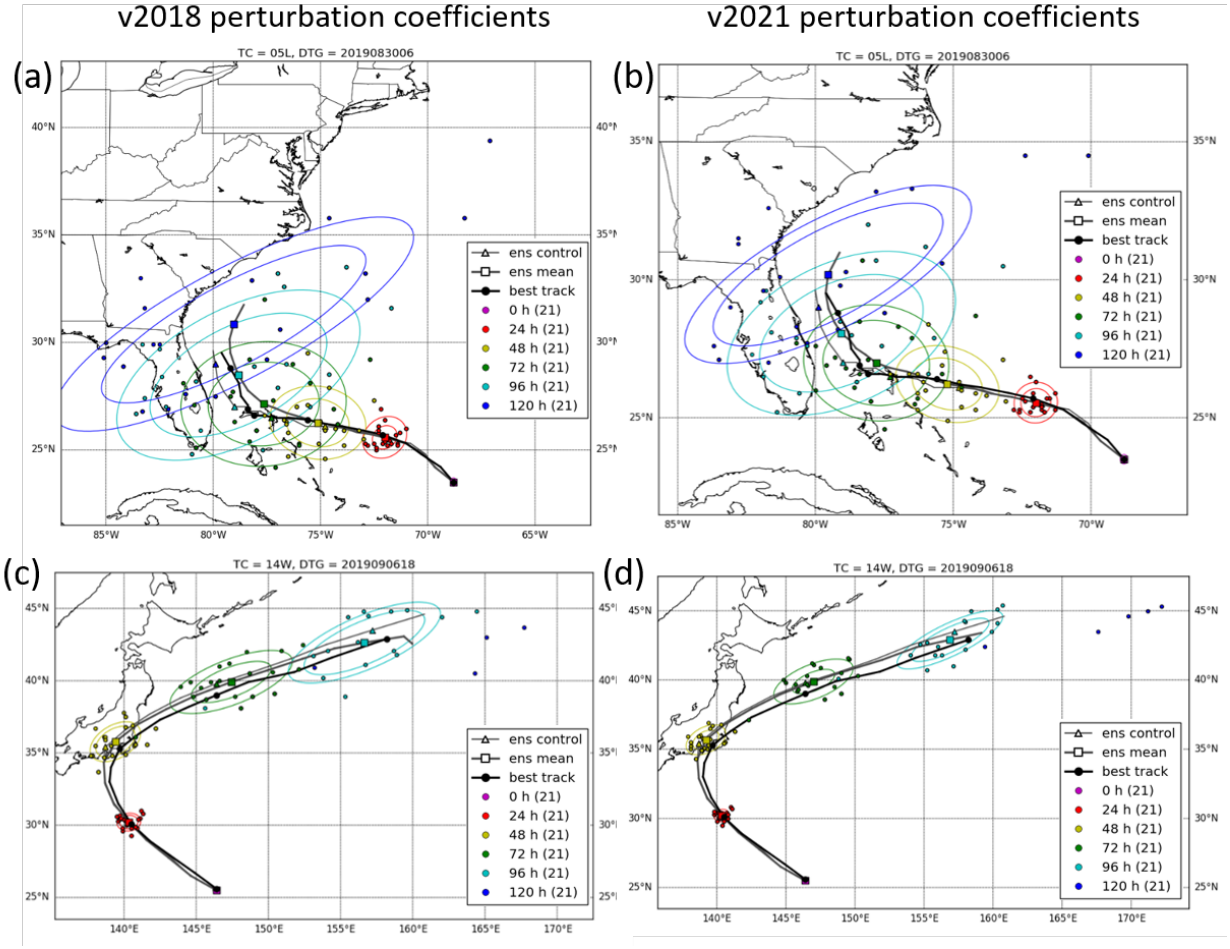


Fig. 1—Track forecast for (a,b) Hurricane Dorian from 0600 UTC 30 August 2019 and (c,d) Typhoon Faxai from 1800 UTC 06 September 2019, using the (a,c) v2018 perturbation coefficients for fcp_pert_ic and fcp_pert_bc, and using the (b,d) v2021 perturbation coefficients

As stated earlier, TCs with initial intensities less than or equal to 55 kt no longer use tcinit to generate an initial vortex. Because tcinit is also used to generate the vortex perturbations, these vortex perturbations are no longer used for TCs of these weak initial intensities. As will be demonstrated in this report, this change improves the intensity mean absolute error (MAE) and RI skill scores throughout the 120 h forecast while only sacrificing a small amount of intensity spread for the initial 24 h of the forecast.

All Python scripts were updated from Python 2 to Python 3 format. This change allowed for an update to the graphics suite from the Basemap Python package to Cartopy. Cartopy features more advanced

plotting capabilities and produces aesthetically cleaner graphics, which should be well-received by our end users. See Fig. 2 for an example using the old and new graphics suite from Hurricane Laura (2020).

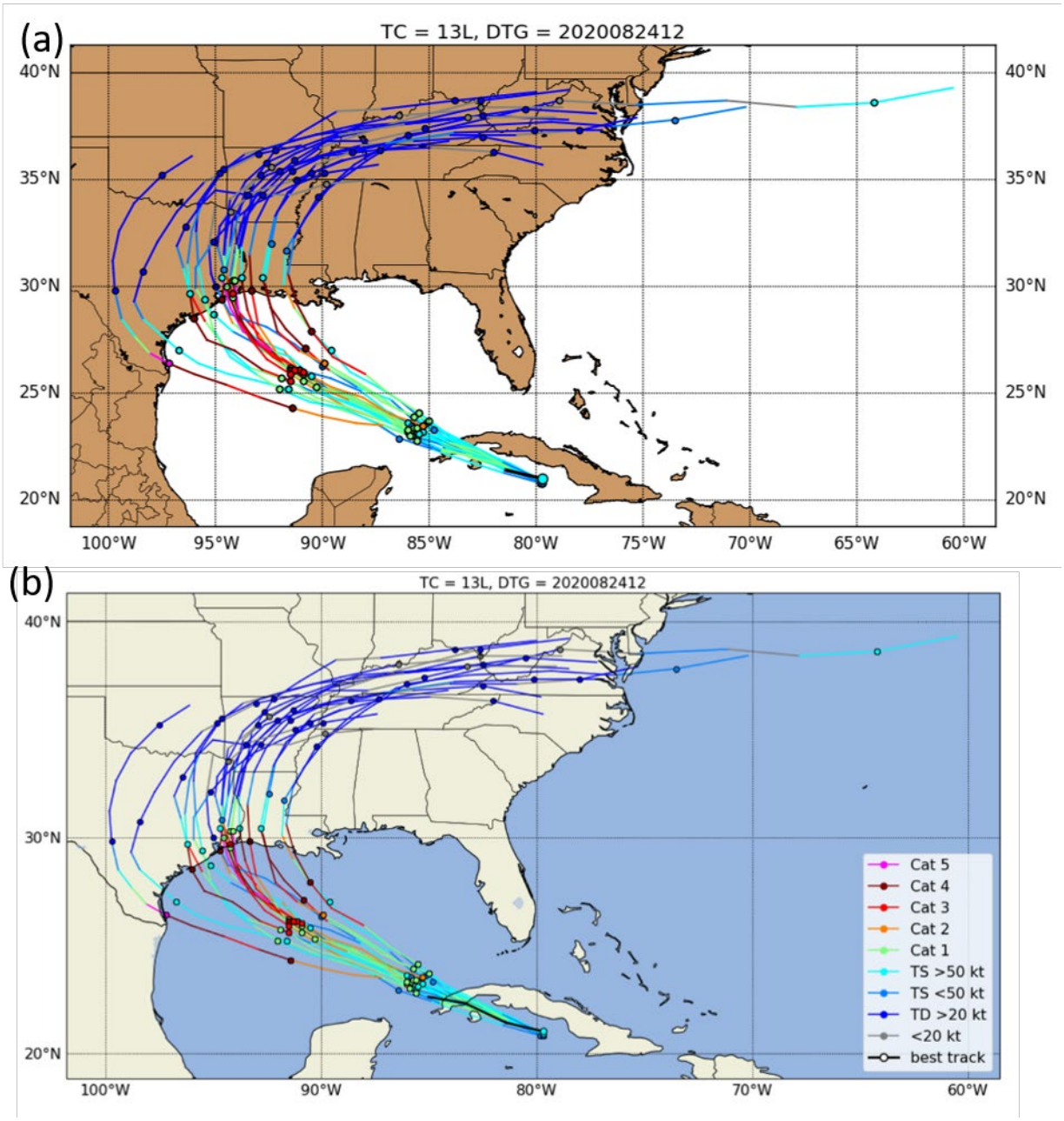


Fig. 2—COAMPS-TC ensemble track color-coded by intensity forecast for Hurricane Laura from 1200 UTC 24 August 2020, plotted using (a) the original graphics suite with Python 2 and Basemap, and (b) the revised graphics suite with Python 3 and Cartopy

Additional model configuration changes include a grid configuration change that allows grids 2 and 3 to move into the grid 1 blendzone. The primary impact of this change is that it allows the TC to be tracked several degrees latitude further north, which is particularly important for recurving Western North Pacific and North Atlantic cyclones. The code was also modified to initialize using the new format FNMOC

"tcvitals" file, replacing the old tcbog.gfdn. The ensemble has also been updated to allow it to run Invests in addition to numbered TCs. This modification is particularly noteworthy, as Invests are occasionally ranked first on the JTWC sorted storm priority list. This occurs when a formative TC is expected to rapidly develop into a TC and intensify near land, particularly for systems in the Western North Pacific near Guam, Okinawa, Taiwan, or the Philippines.

3. ENSEMBLE CONTROL MEMBER VERIFICATION

3.1 Overview of Cases

In this section, we will compare the performance of the unperturbed ensemble control member from v2021 against the v2018 control run. For the ensemble control member, the testing methodology is the same as that used in testing v2021 deterministic COAMPS-TC (transitioned on 30 March 2021). Note that for the full 11-member ensemble a different, smaller sample is used (as documented in the next section). Using GFS initial and boundary conditions (i.e. CTCX), the v2021 ensemble control was compared to the v2018 ensemble control for retrospective cases in the Atlantic, Eastern North Pacific, and Western North Pacific basins from 2018–2020. GFS model analyses and forecast fields are drawn from the real-time operational runs, not retrospective runs. For a particular TC in the sample, forecasts are run every 24 h. As such, forecasts in the sample are considered quasi-independent.

The sample used in this part of the study appears in Table 2. Considering all basins, the full sample is 412 cases from 85 TCs. The Western North Pacific sample consists of 181 cases from 37 TCs, from which nearly all the longer-lived TCs in 2018–2020 are included in the sample. The Atlantic sample consists of 147 cases from 27 TCs, from which nearly all longer-lived TCs from 2018–2020 are included in the sample, including 15 TCs from the historic 2020 season. The Eastern North Pacific sample consists of 79 cases from 16 TCs in 2018–2020, including all longer-lived TCs from 2019 and 2020. Lastly, to ensure robust performance worldwide, a TC is tested in every regional grid; this includes one TC from the Central North Pacific, two from the North Indian Ocean, and two from the Southern Hemisphere.

Table 2—Case List for the Ensemble Control Retro Sample

WestPac

Storm	Cases
wp092018	3
wp102018	5
wp122018	3
wp152018	9
wp172018	5
wp222018	7
wp262018	8
wp282018	7
wp302018	7
wp312018	9
wp332018	4
wp092019	4
wp102019	6
wp112019	9
wp142019	5
wp152019	4
wp192019	4
wp202019	5
wp212019	3
wp222019	5
wp242019	5
wp262019	3
wp272019	3
wp292019	7
wp012020	3
wp032020	2
wp092020	3
wp102020	4
wp112020	5
wp142020	3
wp152020	2
wp162020	6
wp192020	5
wp212020	3
wp222020	8
wp232020	4
wp252020	3
Total	**181**

Atlantic

Storm	Cases
al032018	4
al062018	13
al082018	7
al092018	6
al102018	4
al142018	3
al052019	11
al082019	5
al092019	4
al102019	6
al122019	3
al132019	7
al032020	4
al082020	3
al092020	4
al112020	3
al132020	7
al142020	4
al172020	7
al182020	6
al192020	4
al202020	8
al222020	3
al262020	3
al272020	4
al282020	3
al292020	11
Total	**147**

EastPac

Storm	Cases
ep142018	9
ep162018	8
ep202018	5
ep022019	1
ep062019	4
ep072019	7
ep112019	4
ep132019	10
ep052020	4
ep082020	7
ep092020	2
ep122020	3
ep142020	3
ep172020	3
ep182020	6
ep192020	3
Total	**79**

Other

Storm	Cases
cp012018	1
io012019	1
io042019	1
sh222020	1
sh252020	1
Total	**5**

3.2 Results for all Basins

First, the MAE for track for all basins is examined. Track MAE is improved in v2021 compared to v2018 beyond 12 h lead time (Fig. 3a), with improvements in the 3% to 5% range (Fig. 3b). These improvements are likely associated with the updated SST cooling parameterization in v2021. The degradations in track MAE at very early lead times in v2021 compared to v2018 are due to GFS downscaling of weak TCs in v2021. However, we found the advantages of GFS downscaling beyond 12 h outweigh this early disadvantage. Note that JTWC never actually uses the 0 h position, but instead uses a corrected version of the 6 h position in formulating the interpolated track forecast, so these initial errors are not operationally relevant.

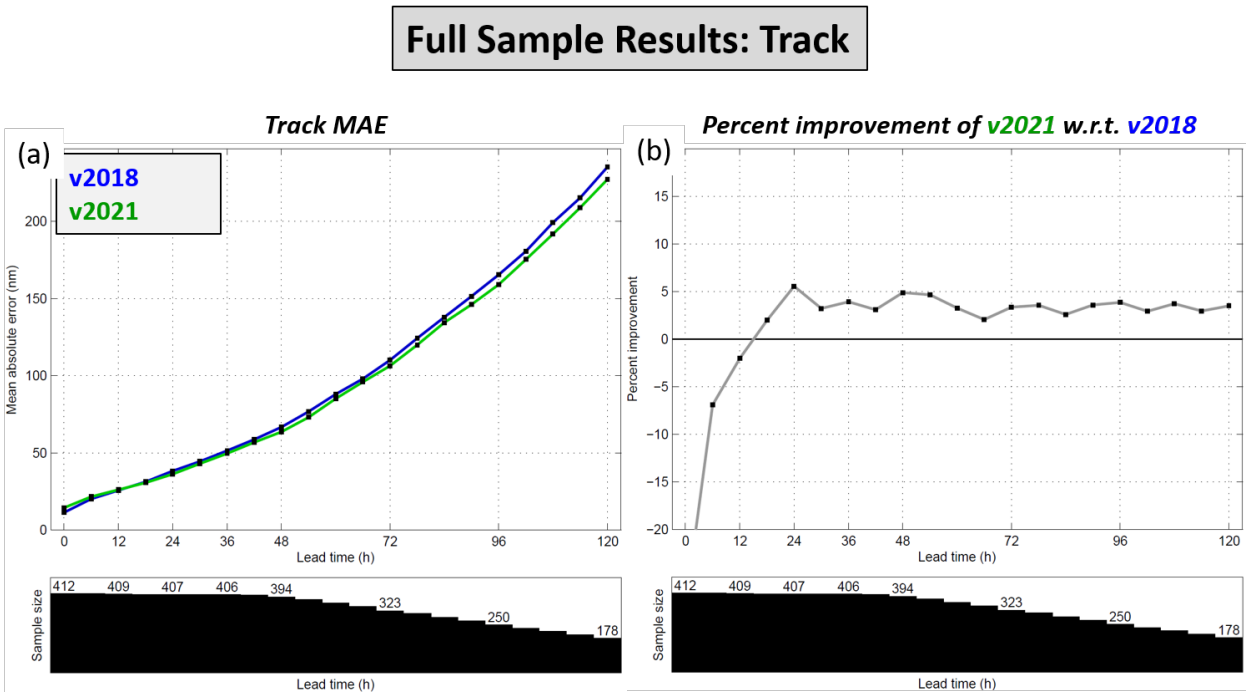


Fig. 3—(a) MAE for track for the unperturbed control member, v2018 (blue) versus v2021 (green), for all basins; and (b) percent improvement for track of v2021 compared to v2018. The number of cases in the sample at each forecast lead time are indicated below each panel.

Intensity MAE is improved in v2021 compared to v2018 for 18–54 h (Fig. 4a) by up to 9% (Fig. 4b). This is largely due to GFS downscaling of initially weak storms in v2021. Intensity MAE results are neutral at later lead times. At very early lead times, there is degradation of intensity MAE in v2021 compared to v2018, also due to GFS downscaling. In terms of ME for intensity (dashed lines), v2021 is superior to v2018 with a substantial reduction in the weak bias characteristic of v2018 from 12–96 h.

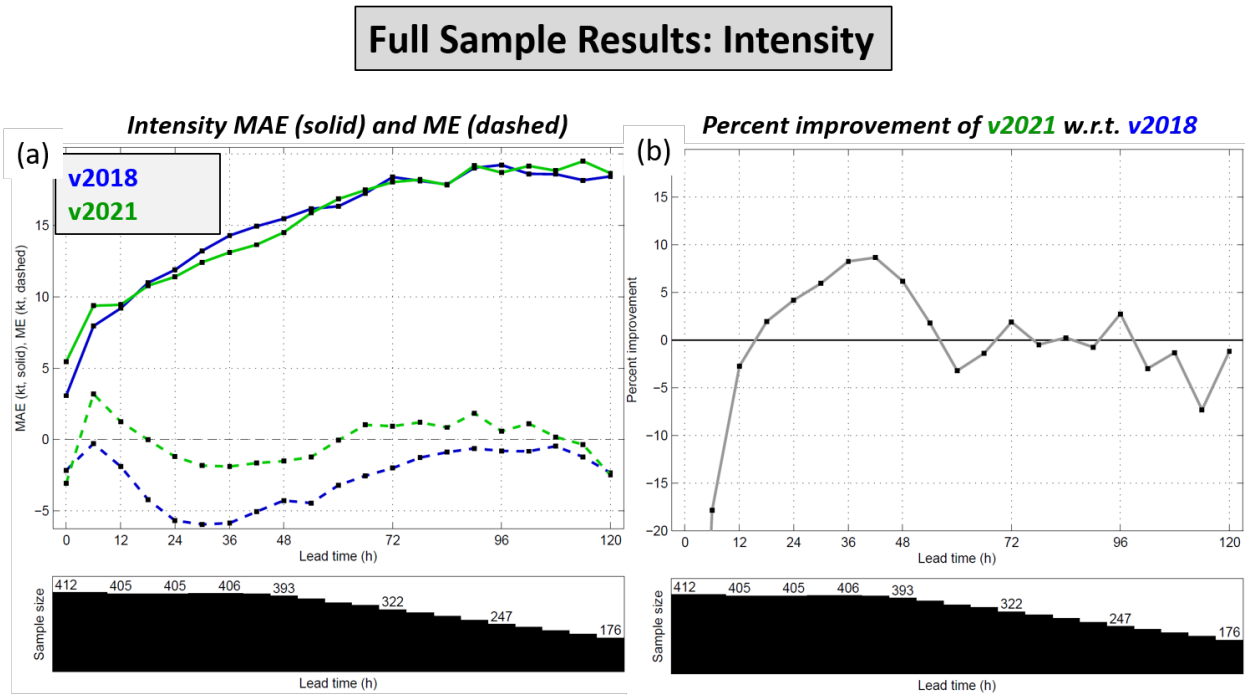
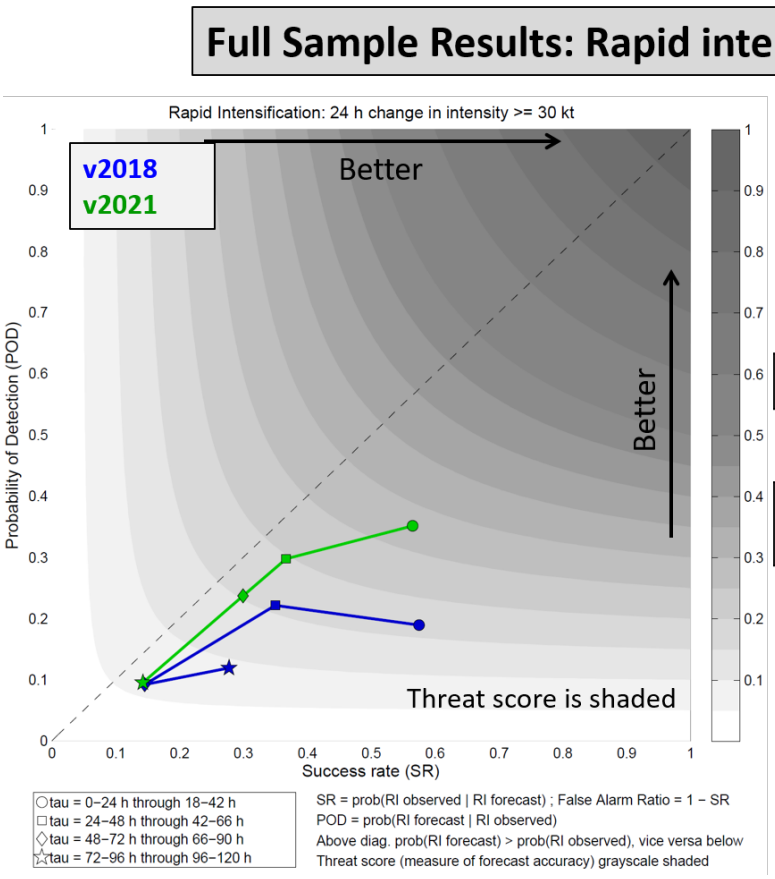


Fig. 4: (a) MAE (solid lines) and ME (dashed lines) for intensity for the unperturbed control member; v2018 (blue) versus v2021 (green) for all basins; and (b) percent improvement for intensity of v2021 compared to v2018. The number of cases in the sample at each forecast lead time are indicated below each panel.

Next, we examine the RI statistics for the two versions of the ensemble control member (Fig. 5). RI is notoriously difficult to predict, and even relatively minor improvements in our ability to predict RI are noteworthy. In terms of event-based prediction of RI (using a 30 kt threshold over a 24 h time window), accuracy is vastly improved in v2021 compared to v2018 (see threat scores in Fig. 5). This is especially true at earlier lead times. RI relative frequency is substantially increased in v2021 versus v2018 (particularly at the earliest lead times) and is closer to the observed relative frequency at all lead times. These improvements to RI prediction in v2021 are due primarily to GFS downscaling for initially weak TCs, updates to the drag coefficient, and updates to the SST cooling parameterization.



Threat Score

	0 - 24 h	24 - 48 h	48 - 72 h	72 - 96 h
v2018	0.17	0.16	0.06	0.09
v2021	0.28	0.20	0.15	0.06

RI relative frequency

	0 - 24 h	24 - 48 h	48 - 72 h	72 - 96 h
Observed	13.7%	10.7%	6.4%	3.9%
v2018	4.5%	6.8%	4.1%	1.7%
v2021	8.5%	8.7%	5.1%	2.6%

Fig. 5—RI statistics for the unperturbed control member, comparing v2018 (blue) to v2021 (green) at various lead times for all basins. The x-axis indicates success rate (SR), the y-axis indicates probability of detection (POD), and the threat score is shaded. Tables indicating the threat score and the RI relative frequency are also included.

The intensity forecast relative frequency distribution is examined to ensure that the model is not over-producing or under-producing TCs of a particular intensity (Fig. 6a), and the pressure-wind relationship is examined as a check that the minimum central pressure for a given intensity approximately matches those observed in nature (Fig. 6b). For both of these distributions, v2018 and v2021 are compared against the NHC and JTWC best track. The intensity distribution shows that v2021 more accurately simulates the relative frequency of intense TCs. This can be attributed to the updated drag coefficient. The pressure-wind relationship is also improved in v2021 compared to v2018, again due to the updated drag coefficient.

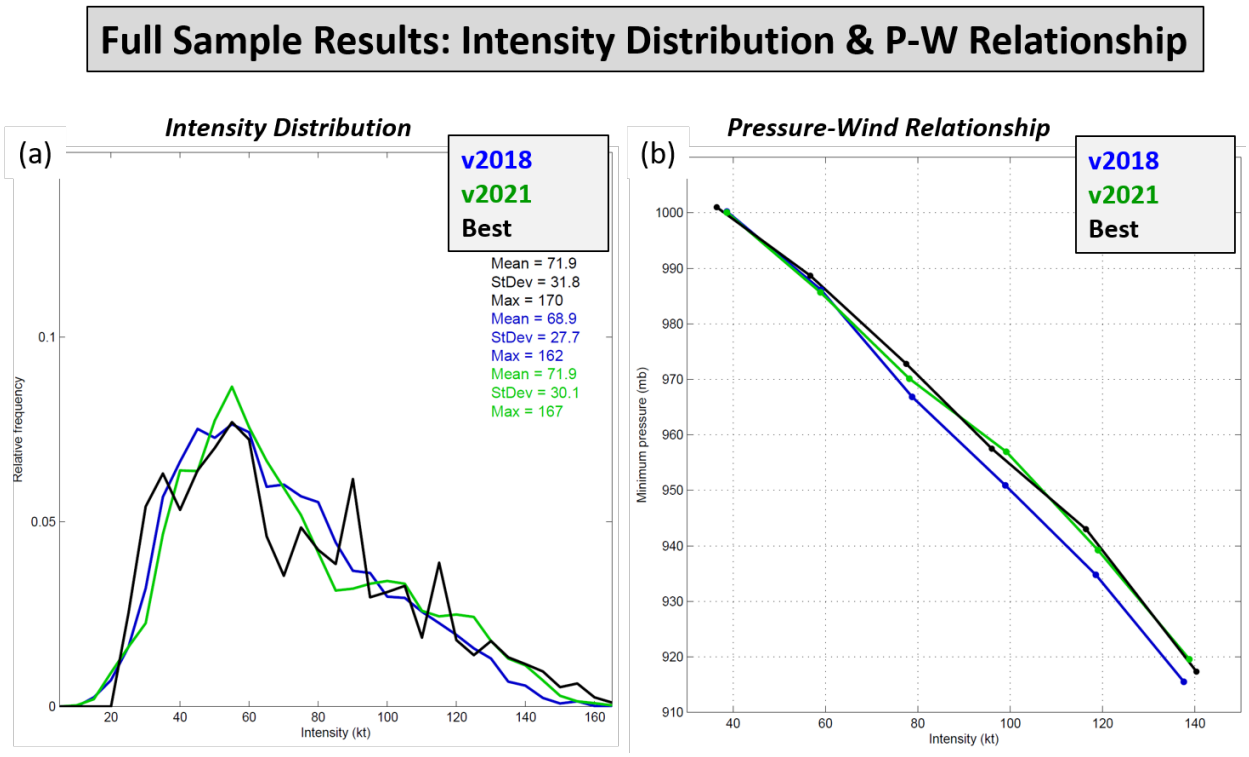


Fig. 6—(a) Intensity relative frequency distribution for v2018 (blue) and v2021 (green) for the unperturbed control member at all forecast lead times, for all basins, against the NHC and JTWC best track (black); and (b) the pressure-wind relationship for all forecast lead times, for all basins, against the NHC and JTWC best track (black).

Wind radii are also examined. The radius of 34 kt wind (R34) MAE and ME are improved at all lead times in v2021 compared to v2018 (Fig. 7a), as is the conditional mean R34 (Fig. 7b). These improvements can be attributed to updates to the SST cooling parameterization and drag coefficient, improvements to *tcinit* for initially strong storms, and the introduction of the Tiedtke shallow and mid-level cumulus on grids one and two. As is the case for R34, MAE and ME for R50 and R64 are improved at all lead times in v2021 compared to v2018, as is the conditional mean R50 and R64 (not shown). The conditional mean radius of maximum wind (Rmax) is mostly similar between v2021 and v2018, but with the higher mean intensity in v2021 compared to v2018, v2021 generally scores better in terms of MAE and ME.

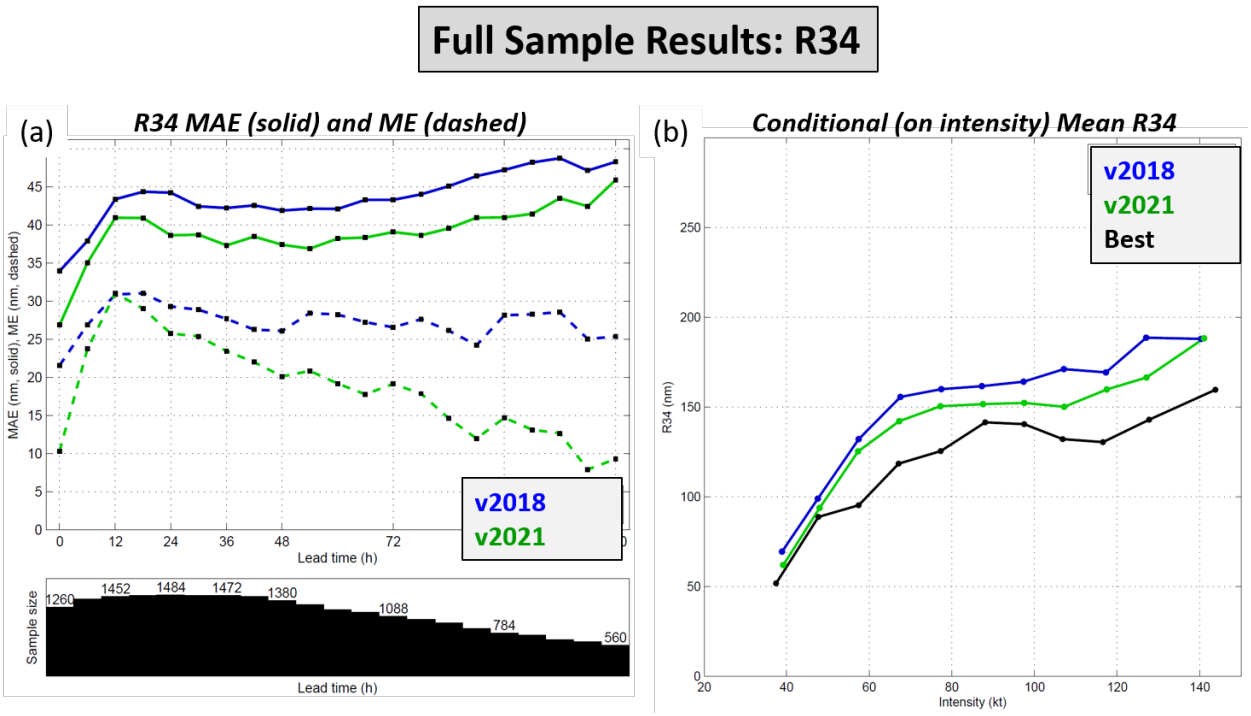


Fig. 7—(a) MAE (solid lines) and ME (dashed lines) for the radius of 34 kt wind (R34; nmi) for the unperturbed control member, v2018 (blue) versus v2021 (green), for all basins; and (b) R34 as a function of TC intensity for v2018 (blue), v2021 (green), and NHC and JTWC best track (black), for all basins. The number of cases in the sample at each forecast lead time are indicated below the left panel.

3.3 West Pacific Results

For brevity, this report does not present results for every ocean basin individually. In general, the results do not vary dramatically from one basin to the next. However, due to its significance to JTWC and the 7th Fleet, this report presents a few results just from the West Pacific subset of the sample.

In the West Pacific, track MAE is substantially improved in v2021 compared to v2018 beyond 6 h lead time, with improvements even larger than those observed in the full sample—generally in the 5% to 10% range (Fig. 8a). Further improvements will be realized once we update the West Pacific outer grid to be consistent with the v2021 deterministic configuration in a future upgrade. Note that the degradations in track MAE at very early lead times in v2021 versus v2018 are due to GFS downscaling of weak TCs in v2021. The northeast track bias is also reduced in v2021 versus v2018, which contributes to the improvement in track MAE (Fig. 8b).

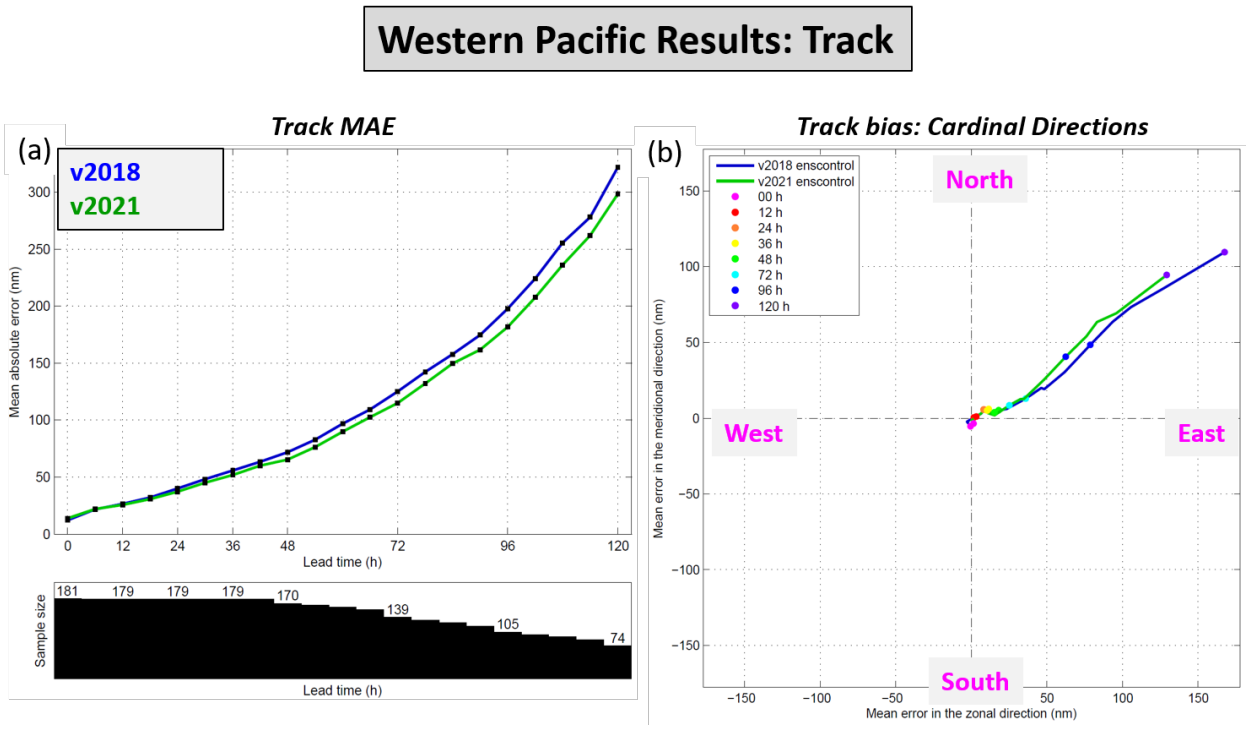


Fig. 8—(a) MAE for track for the unperturbed control member, v2018 (blue) versus v2021 (green), for the West Pacific; and (b) change in track bias in the West Pacific between v2018 (blue) and v2021 (green). The number of cases in the sample at each forecast lead time are indicated below the left panel.

Regarding intensity, MAE is improved in v2021 compared to v2018 for most lead times in the West Pacific (Fig. 9a). The largest MAE improvements (of up to 16%) occur around 36 h lead time and can be attributed mostly to GFS downscaling of initially weak storms in v2021 (Fig. 9b). Intensity MAE is slightly degraded at the very latest lead times in v2021, and there is degradation at very early lead times as well. This is mostly attributed to GFS downscaling. In terms of intensity ME, v2021 is superior to v2018 at early and middle lead times.

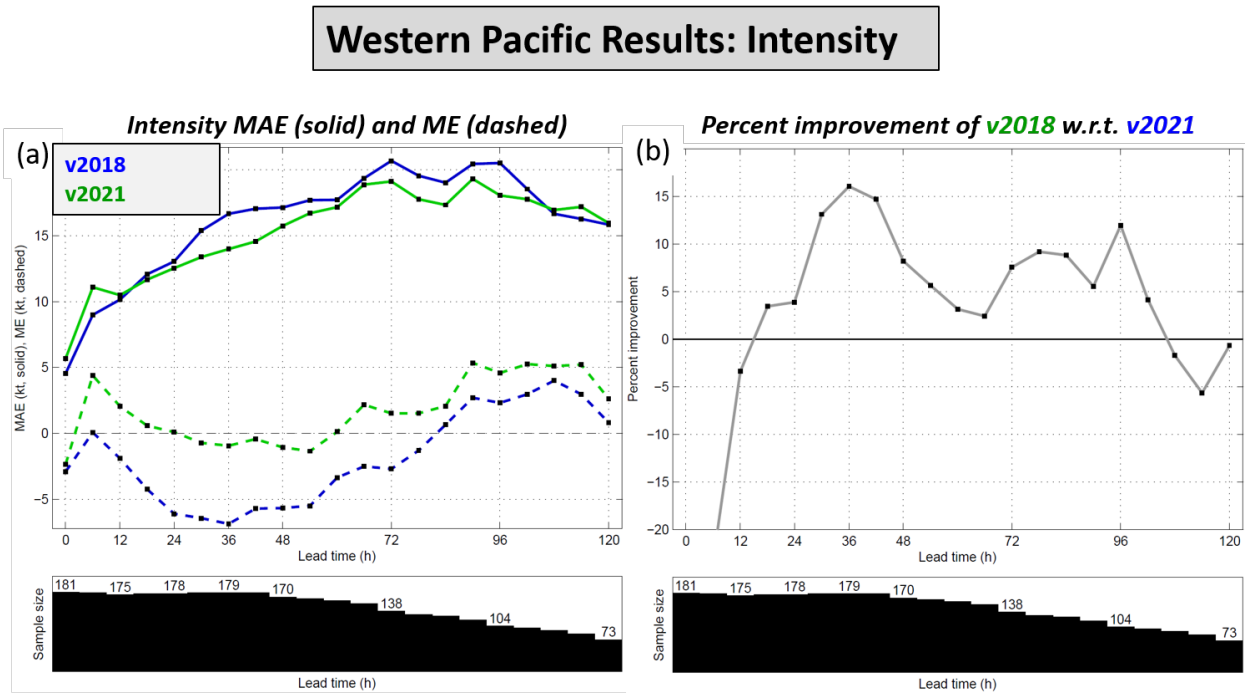
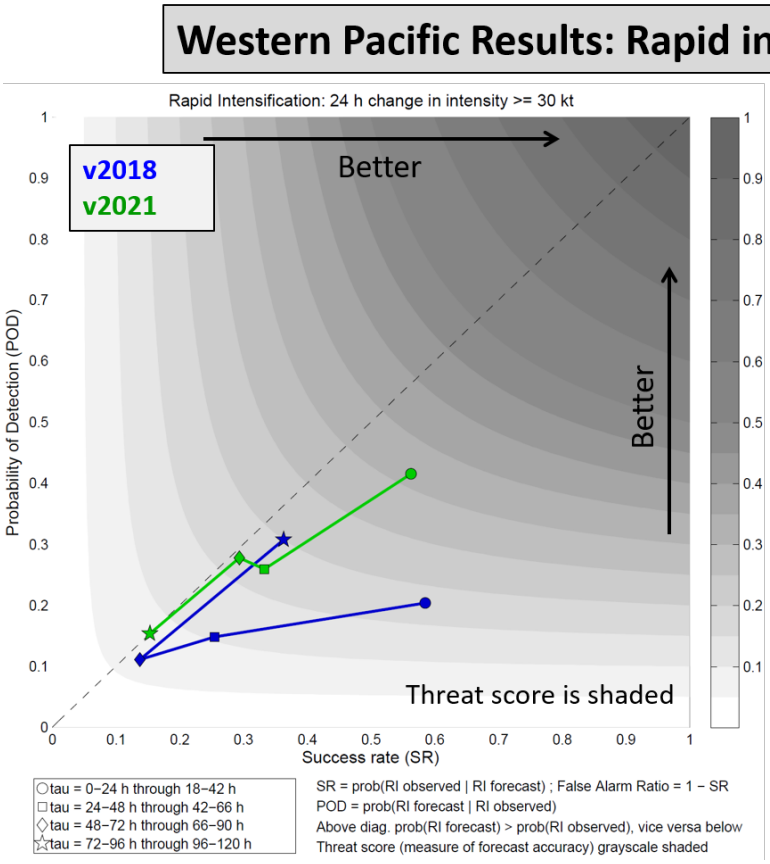


Fig. 9—(a) MAE (solid lines) and ME (dashed) for intensity for the unperturbed control member; v2018 (blue) versus v2021 (green), for the West Pacific; and (b) percent improvement for intensity of v2021 compared to v2018. The number of cases in the sample at each forecast lead time are indicated below each panel.

In event-based prediction of RI (using a 30 kt threshold over a 24 h time window) for the West Pacific, accuracy is vastly improved in v2021 compared to v2018 (excluding the latest lead times (Fig. 10)). RI relative frequency is substantially increased in v2021 versus v2018 (particularly at the earliest lead times) and is closer to the observed relative frequency at all lead times. These improvements to RI prediction in v2021 are due primarily to GFS downscaling for initially weak TCs, as well as updates to the drag coefficient and SST cooling parameterization.



Threat Score

	0 - 24 h	24 - 48 h	48 - 72 h	72 - 96 h
v2018	0.18	0.10	0.07	0.20
v2021	0.31	0.17	0.17	0.08

RI relative frequency

	0 - 24 h	24 - 48 h	48 - 72 h	72 - 96 h
Observed	16.6%	12.5%	7.0%	2.9%
v2018	5.8%	7.2%	5.7%	2.4%
v2021	12.3%	9.7%	6.7%	2.9%

Fig. 10—RI statistics for the unperturbed control member, comparing v2018 (blue) to v2021 (green) at various lead times, for the West Pacific. The x-axis indicates success rate (SR), the y-axis indicates probability of detection (POD), and the threat score is shaded. Tables indicating the threat score and the RI relative frequency are also included.

Improvements to wind radii in the West Pacific are consistent with improvements observed in the full sample (see Section 3.2). Improvements to track and intensity MAE and ME in the Atlantic and East Pacific are overall similar to the full sample, although with smaller magnitude improvements than what was observed in the West Pacific.

4. FULL 11-MEMBER ENSEMBLE FORECASTS

4.1 Overview of Cases

The testing methodology for the full 11-member retrospective runs is similar to that employed for the ensemble control member; however, due to computational cost constraints, we used a 180-case subset of the 412-case retrospective sample used for the ensemble control member. For a particular TC in the sample, forecasts are run every 48 h. As such, forecasts in the sample can reasonably be considered independent. The retrospective testing sample for the full 11-member ensemble is the same as for the unperturbed control (Table 2), except that it is subsampled every 48 h instead of every 24 h, and we excluded some forecast cases that did not have an extant TC in the corresponding best track to verify against beyond 48 h lead time. Forecasts for 9 of 11 ensemble members must exist for a particular case/lead time to be included in the validation sample. Using fewer than 9 members can produce a non-representative ensemble mean. As an example, if half the members make landfall and dissipate after 48 h while half the members miss land and survive through 120 h, the ensemble mean will jump after 48 h to match the mean

of just the offshore members. If the forecaster only uses the mean, this will suggest an unrealistically high chance of remaining offshore at 72 h.

4.2 Deterministic Validation of the Ensemble Mean

First, we examined track MAE comparing v2021 to v2018 for the mean of the full 11-member ensemble, using the full sample across all basins. It was found that track MAE is improved in v2021 compared to v2018 beyond 18 h lead time (Fig. 11a), with improvements in the 2-5% range from 24 to 72 h and in the 5-8% range at later lead times (Fig. 11b). The degradations in track MAE at very early lead times in v2021 versus v2018 are due to GFS downscaling of weak TCs in v2021. At later leads, the reduced frequency of *outlier* member tracks in v2021 likely contributes to its improved performance.

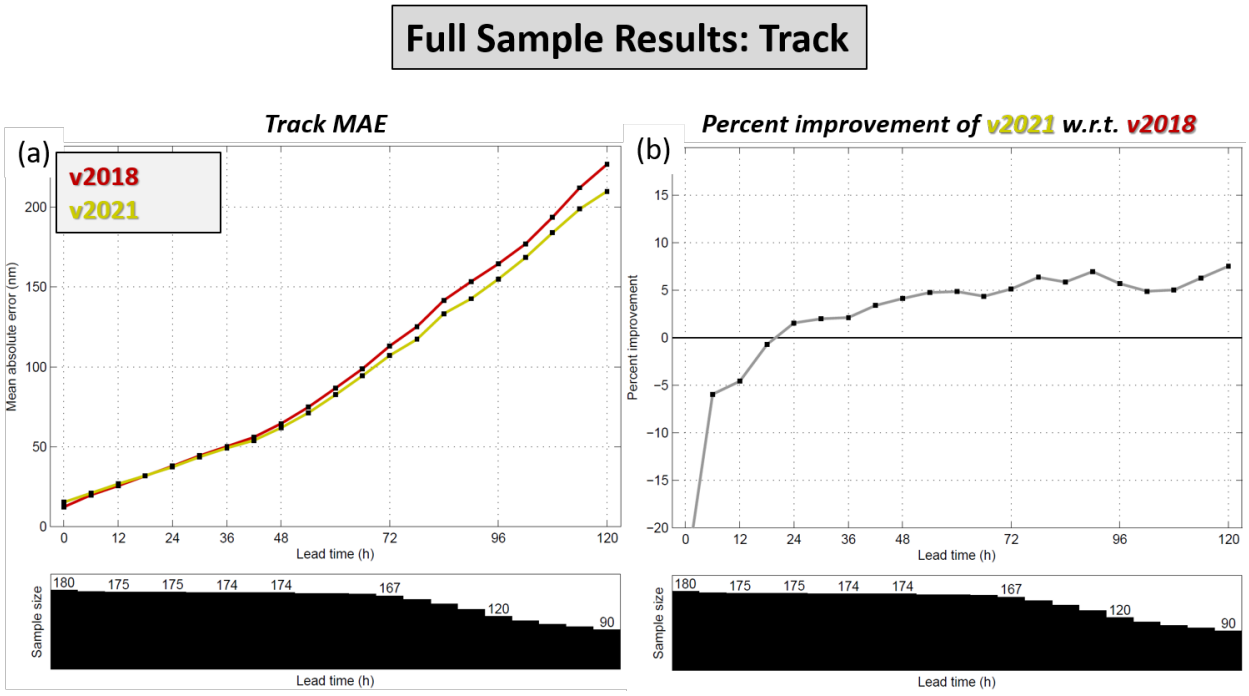


Fig. 11—(a) MAE for track for the mean of the full 11-member ensemble, v2018 (red) versus v2021 (yellow) for all basins; and (b) percent improvement for track of v2021 compated to v2018. The number of cases in the sample at each forecast lead time are indicated below each panel.

In terms of intensity performance, MAE is improved in v2021 compared to v2018 at all but the earliest lead times (Fig. 12a). The largest improvements of approximately 10% occur around 36 h, driven mostly by improved prediction of initially weak storms due to GFS downscaling in v2021 (Fig. 12b). At very early lead times, there is degradation of intensity MAE in v2021, again due to GFS downscaling. In terms of intensity ME, v2021 is superior to v2018, with a significant reduction in the weak bias that occurred in v2018.

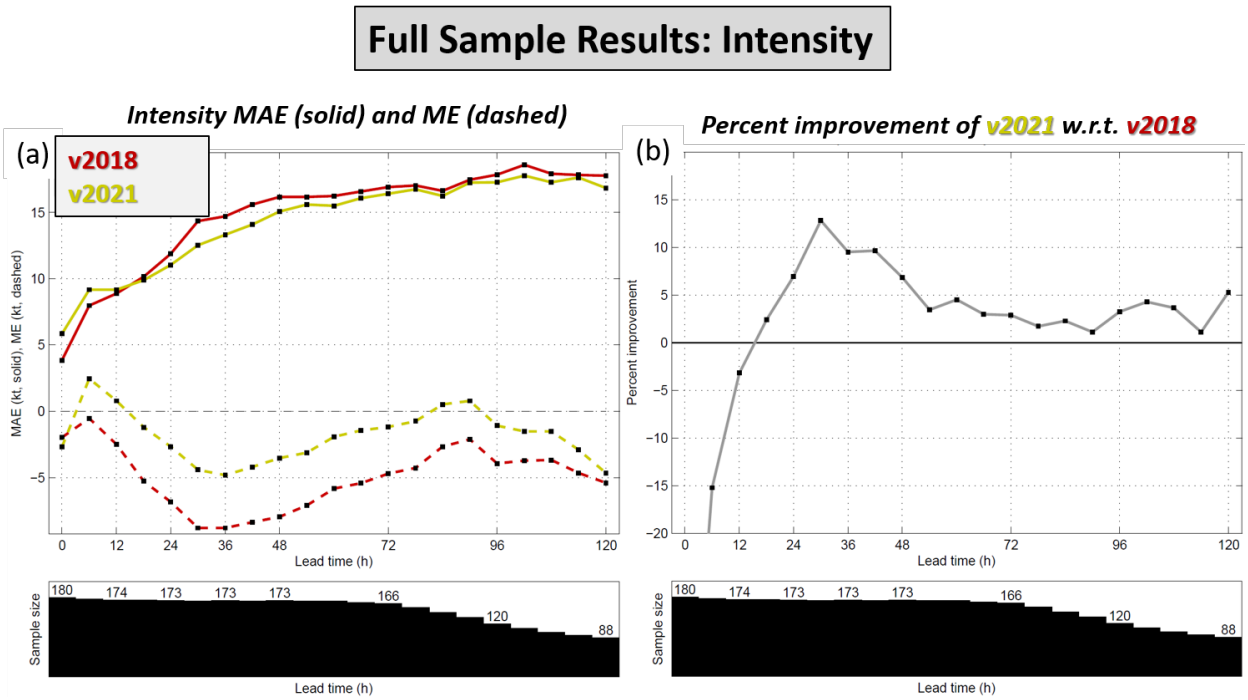
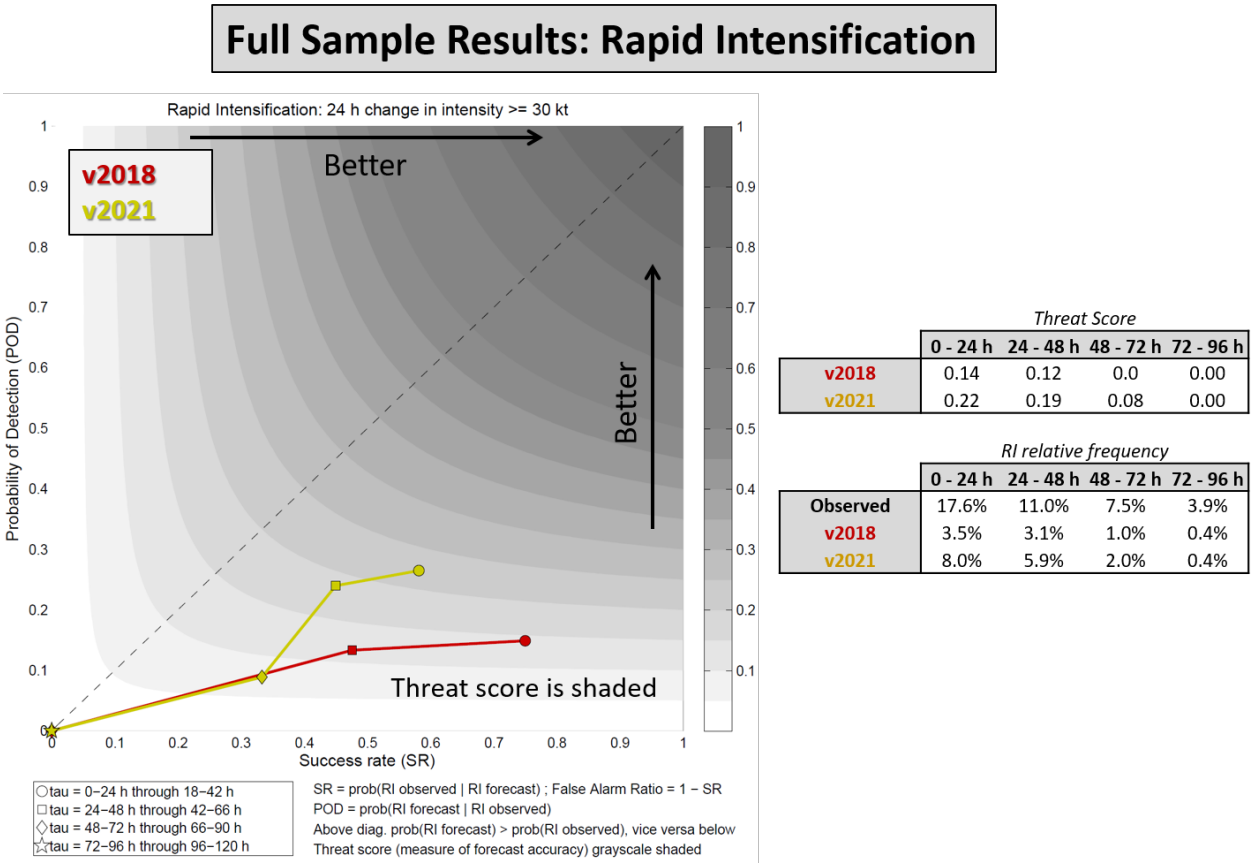


Fig. 12—(a) MAE (solid lines) and ME (dashed) for intensity for the mean of the full 11-member ensemble, v2018 (red) versus v2021 (yellow) for all basins, and (b) percent improvement for intensity of v2021 compared to v2018. The number of cases in the sample at each forecast lead time are indicated below each panel.

For completeness, RI statistics of the ensemble mean are also examined. It must be noted that because the ensemble mean can dampen out large intensity changes that occur at different times across various members, the ensemble mean is not the best ensemble-based tool for assessing the likelihood of RI. With that caveat in mind, RI performance statistics for the ensemble mean signify that the v2021 ensemble members are much more likely to predict RI and are more accurate in predicting RI compared to the v2018 ensemble members (Fig. 13).



Threat Score

	0 - 24 h	24 - 48 h	48 - 72 h	72 - 96 h
v2018	0.14	0.12	0.0	0.00
v2021	0.22	0.19	0.08	0.00

RI relative frequency

	0 - 24 h	24 - 48 h	48 - 72 h	72 - 96 h
Observed	17.6%	11.0%	7.5%	3.9%
v2018	3.5%	3.1%	1.0%	0.4%
v2021	8.0%	5.9%	2.0%	0.4%

Fig. 13—RI statistics for the mean of the full 11-member ensemble, comparing v2018 (red) to v2021 (yellow) at various lead times for all basins. The x-axis indicates success rate (SR), the y-axis indicates probability of detection (POD), and the threat score is shaded. Tables indicating the threat score and the RI relative frequency are also included.

4.3 Probabilistic Verification

One method by which the probabilistic forecast performance of an ensemble may be assessed is by examining the relationship between the average spread of the ensemble about its mean ("spread") and the average error of the ensemble mean ("skill"). If the spread is lower than the ME, then the ensemble is underdispersive (not enough forecast diversity); if the spread is greater than the ME, then the ensemble is overdispersive (too much forecast diversity). The spread-skill score is a metric used to evaluate whether or not the spread in the ensemble is well calibrated with the ME; it is defined as the sample-size weighted, lead-time averaged absolute difference between spread and skill. A spread-skill score of zero is the best possible value.

Here we examine the spread-skill relationship for both track and intensity. Track spread is significantly reduced in v2021 compared to v2018 (Fig. 14) due to the adjustment to the synoptic-scale initial time and lateral boundary perturbation scalars fcp_pert_ic and fcp_pert_bc. The accuracy, or "skill," of the ensemble mean track forecast is also improved in v2021 compared to v2018, but nonetheless the gap between spread and skill is meaningfully wider at all lead times in v2021. However, we would argue that the reduction of unrealistic track scenarios, in which a member or two tracks the TC directly into the subtropical ridge, is a worthwhile tradeoff. We also expect that a forthcoming upgrade to the ensemble in which the grid 1 domains are adjusted, which has already been implemented in v2021 deterministic

COAMPS-TC, will further reduce the ensemble mean track error (without impacting spread) such that the spread-skill score is improved.

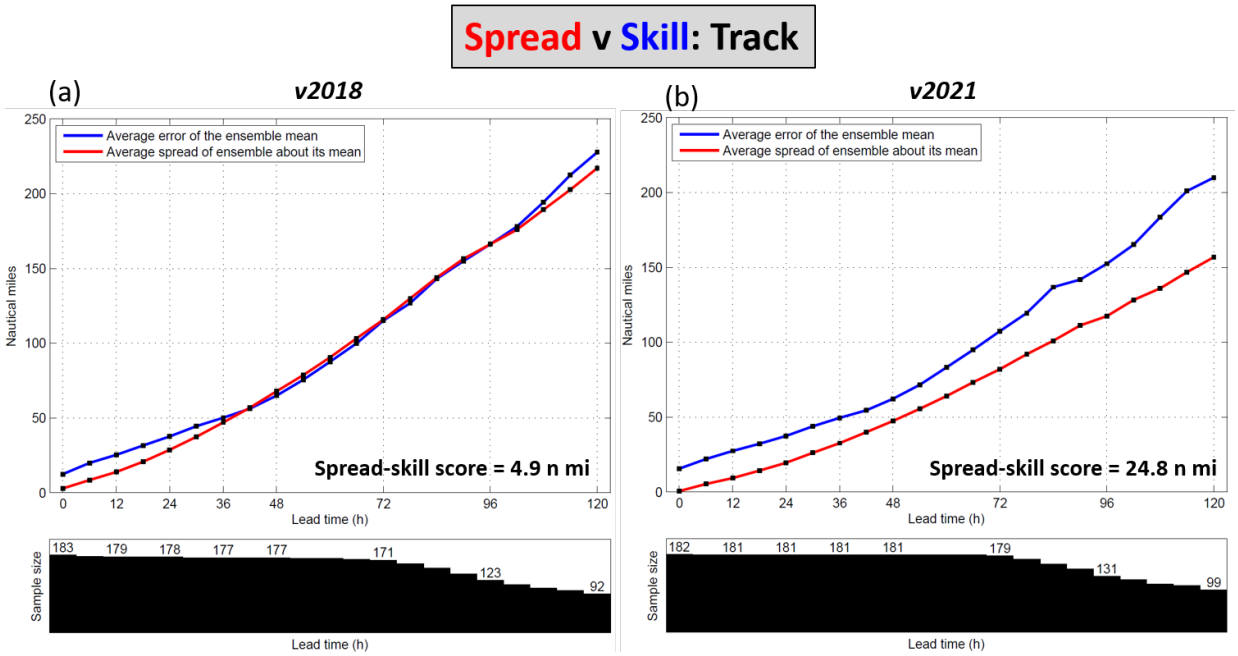


Fig. 14—Spread-skill score, which is a sample-size weighted, lead-time averaged absolute difference between spread and skill, for track (low score = good) for (a) v2018 and (b) v2021. The average error of the ensemble mean (blue) and the average spread of the ensemble about its mean (red) are shown. The number of cases in the sample at each forecast lead time are indicated below each panel.

For intensity, spread tends to be 1.5 to 2.0 kt lower in v2021 compared to v2018 due to the lack of initial time vortex perturbations for weak storms and reduced track diversity (Fig. 15). However, the accuracy or "skill" of the ensemble mean intensity forecast is improved by a similar amount in v2021 compared to v2018, so the gap between spread and skill is largely similar between the two versions with a spread-skill score of 6.2 kt for v2018 and a score of 6.7 kt for v2021.

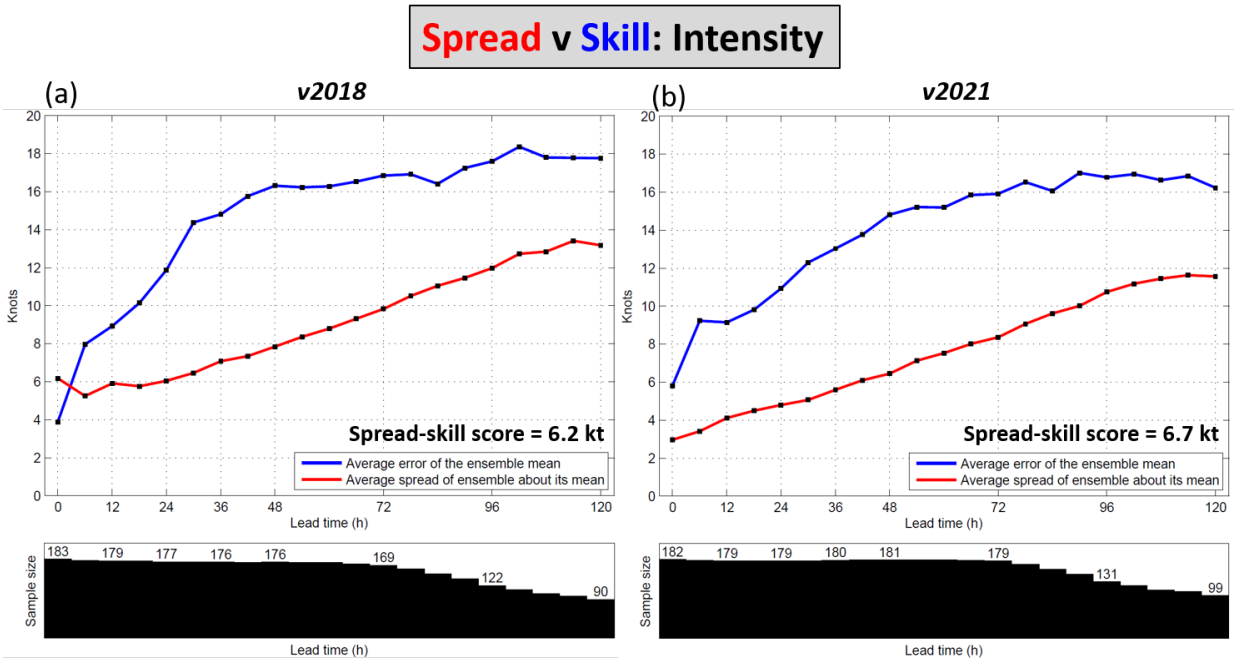


Fig. 15—Spread-skill score, which is a sample-size weighted, lead-time averaged absolute difference between spread and skill, for intensity (low score = good) for (a) v2018 and (b) v2021. The average error of the ensemble mean (blue) and the average spread of the ensemble about its mean (red) are shown. The number of cases in the sample at each forecast lead time are indicated below each panel.

Next, we examine the uncertainty discrimination of the ensemble. While the spread-skill relationship at a given lead time (here, we'll say 48 h) compares the mean spread of all 48 h forecasts to the ME of all 48 h forecasts, our uncertainty discrimination diagnostic takes those 48 h forecasts and bins them into four equally sized quartiles of spread (from least spread to most spread) and computes the ME of each bin. Doing so allows us to assess whether the spread at any given lead time is indicative of the forecast uncertainty at that lead time. In the uncertainty discrimination plots, the red line is a linear least squares fit to the quartile bins. If the slope = 1, then the ensemble is ideally discriminating between high and low uncertainty cases. However, any positive slope indicates some ability to discriminate between high and low uncertainty cases. For the 24 h lead time results, v2018 performs better than v2021 in terms of track uncertainty discrimination (Fig. 16a, b). At later lead times, including at 120 h, v2021 performs better than v2018 (Fig. 16c, d). Considering all lead times, the uncertainty discrimination score (which compares the slope of the least squares fit line to the ideal slope of one) for track indicates slightly better performance in v2018, with a score of 0.40, than for v2021, with a score of 0.46, where a lower score is better.

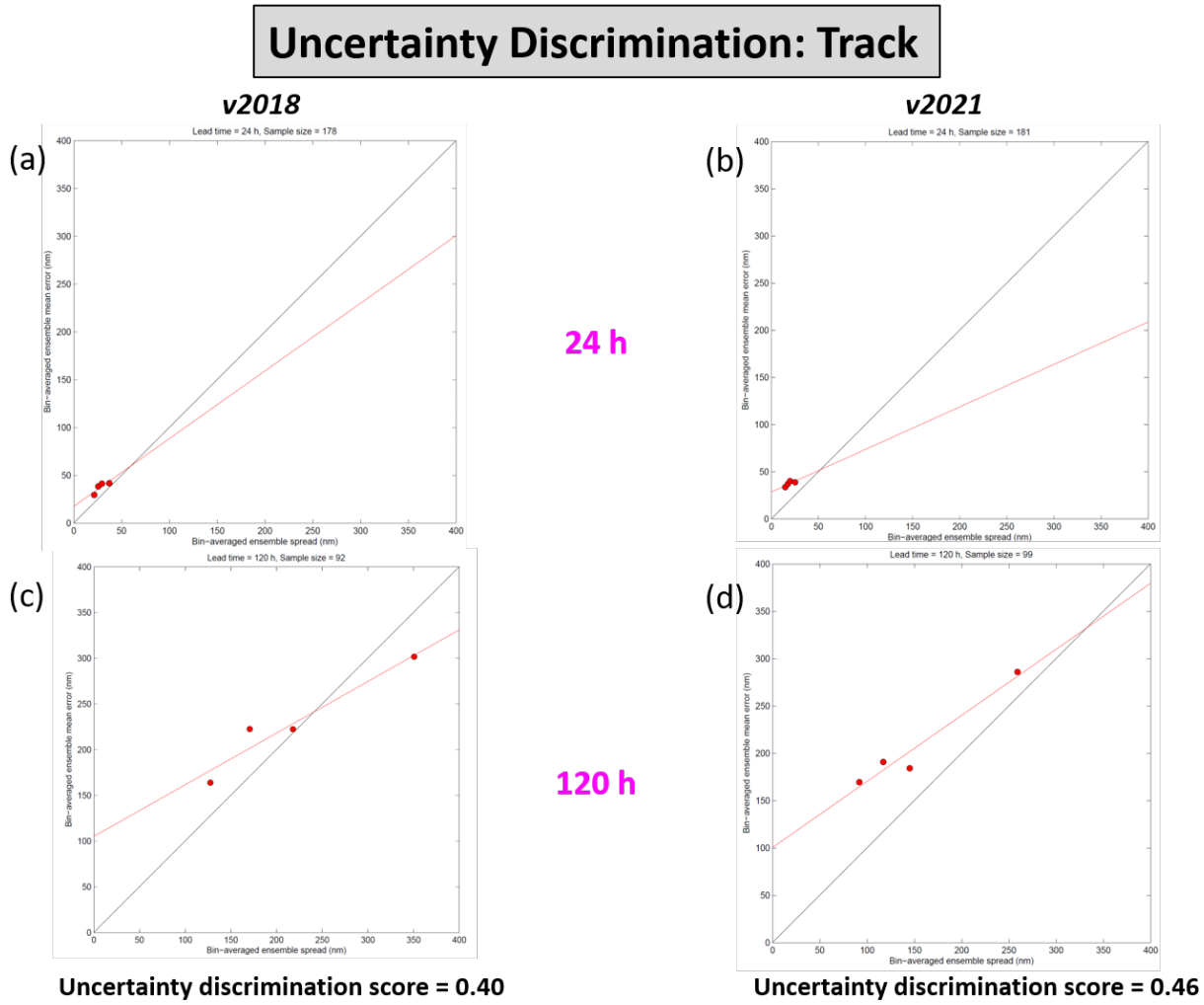


Fig. 16—Uncertainty discrimination for track for (a, c) v2018, and (b, d) v2021 at forecast lead times of (a, b) 24 h and (c, d) 120 h. Each red dot indicates the mean of a quartile bin. The red line is the linear least squares fit, and the grey line is the 1:1 ratio. If the slope = 1, then the ensemble is ideally discriminating between high and low uncertainty cases; however, any positive slope indicates some ability to discriminate between high and low uncertainty cases. The *uncertainty discrimination score* is averaged over all forecast lead times.

Uncertainty discrimination for intensity is also examined. It is found that intensity uncertainty discrimination is improved in v2021 compared to v2018 at 24 h (Fig. 17a, b) and at all lead times except for 120 h (Fig. 17c, d). Considering all lead times, the uncertainty discrimination score indicates overall substantially better performance in v2021 compared to v2018, with respective scores of 0.41 and 0.53.

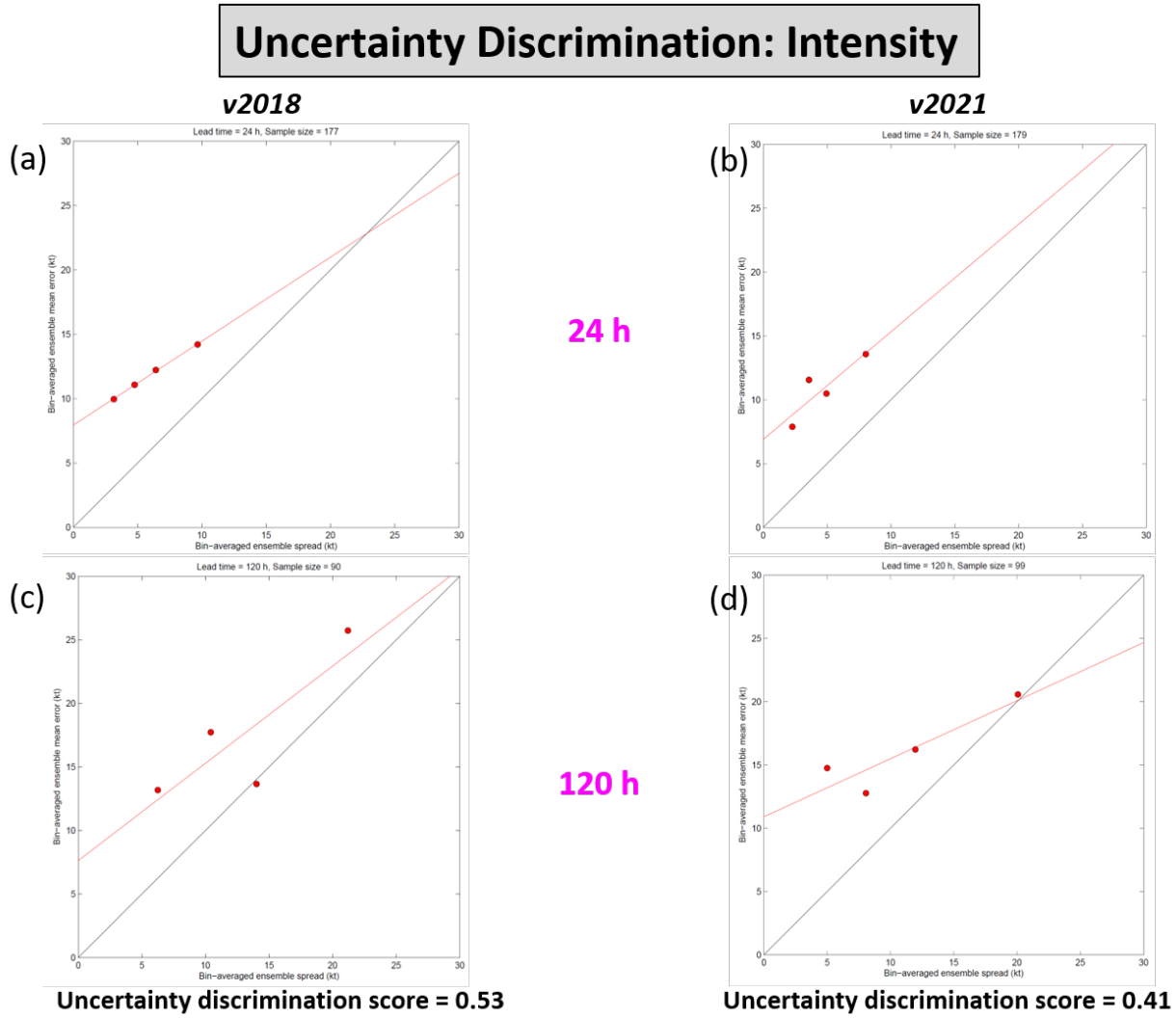


Fig. 17—Uncertainty discrimination for intensity for (a, c) v2018 and (b, d) v2021 at forecast lead times of (a, b) 24 h and (c, d) 120 h. Each red dot indicates the mean of a quartile bin. The red line is the linear least squares fit, and the grey line is the 1:1 ratio. If the slope = 1, then the ensemble is ideally discriminating between high and low uncertainty cases; however, any positive slope indicates some ability to discriminate between high and low uncertainty cases. The *uncertainty discrimination score* is averaged over all forecast lead times.

Rank histograms for intensity are the final of our probabilistic verification metrics examined. The rank histograms for intensity indicate where the verifying intensity falls in the spectrum of predicted intensity, where each member is sorted from lowest to greatest intensity. A flat rank histogram distribution would indicate that it is equally probable that the verifying intensity falls between any two ensemble members (bins 2–11) or outside the ensemble envelope on either end (bins 1 and 12). The results indicate at early lead times, 6–24 h, the verifying intensity is greater than any predicted intensity for both v2018 and v2021 approximately 24% of the time (Fig. 18a, b). However, the v2021 histogram is more evenly weighted, with the verifying intensity also much more likely to be weaker than the entire forecast distribution in v2021 than in v2018. At later lead times, from 102–120 h, the results are actually better and the rank histograms are flatter (Fig. 18c, d). The verifying intensity being above (below) the entire ensemble forecast distribution only approximately 17% (14%) of the time in v2021. Considering all lead times, v2021

has somewhat more relative frequency in the outermost bins compared to v2018 (which is bad), but the rank histogram is more symmetric, indicating better centering of the distribution (which is good).

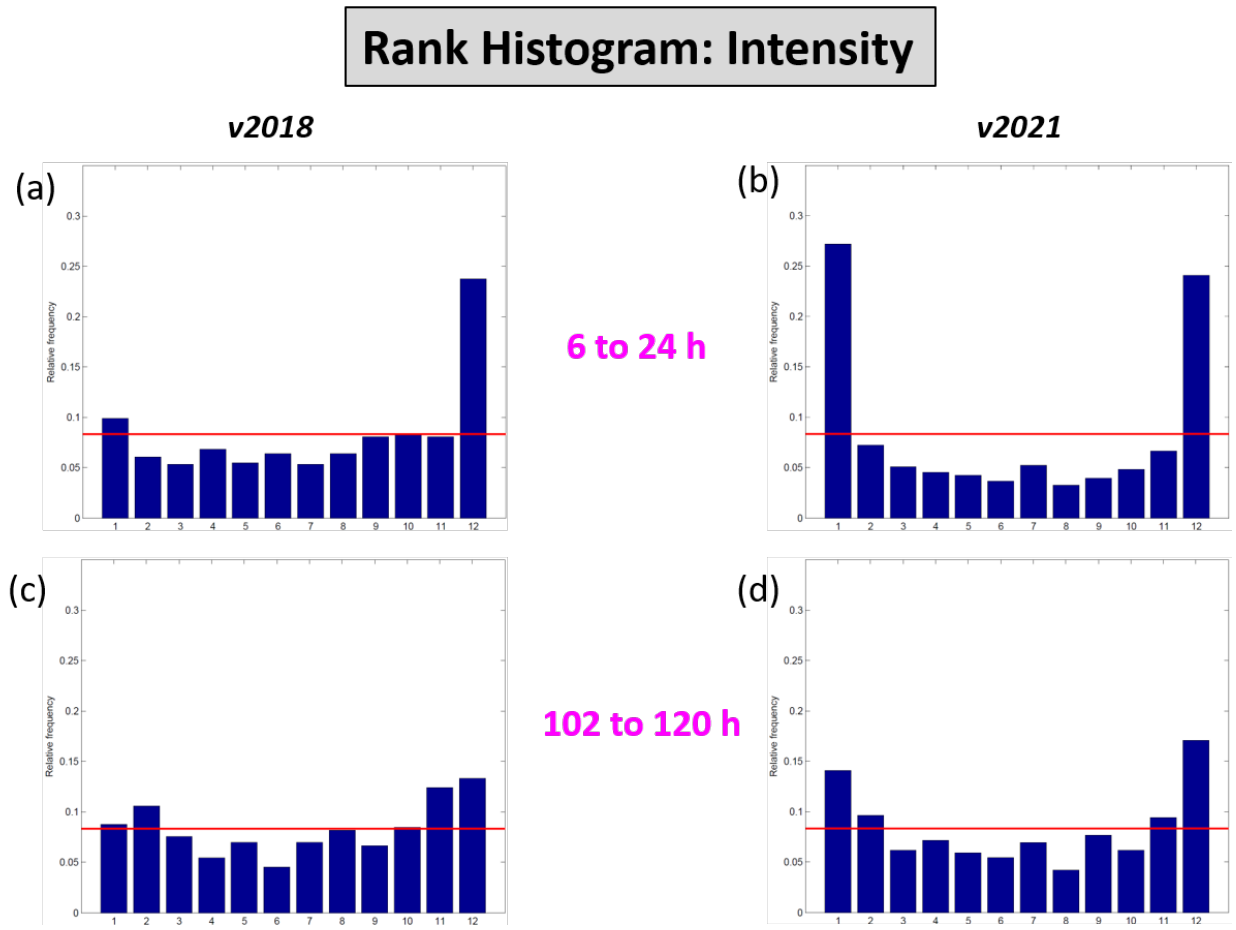


Fig. 18—Rank histogram for intensity for (a, c) v2018 and (b, d) v2021, for (a, b) 6–24 h and (c, d) 102–120 h forecast lead times. In these plots, we are looking for a flat rank histogram (blue bars along red line). This would indicate that it is equally probable that the verifying intensity falls between any two ensemble members (bins 2–11) or outside the ensemble envelope on either end (bins 1 & 12).

5. SUMMARY AND CONCLUSIONS

Overall, the results presented in this VTR indicate that the v2021 COAMPS-TC ensemble is a substantial improvement compared to v2018. To restate, this transition includes: adjusted synoptic-scale initial time and lateral boundary perturbation magnitudes; updates to the shallow cumulus parameterization on grids 1 and 2; modifications to *tcinit* to produce smaller, more realistic TCs, the implementation of graupel-radiation interaction; initialization off of 0.25 deg GFS (versus 0.50 deg GFS, as is presently done); GFS downscaling for weak TCs to produce a more realistic, less symmetric initial vortex for weaker storms; an updated surface drag coefficient; an improved 1-dimensional sea surface temperature cooling parameterization; changes to the interaction between grids 2 and 3 with the grid 1 blend zone; and increased diffusion in the first 6 h of the forecast for weak TCs. The ensemble has also been modified such that it can now run Invests, which are often high on JTWC's priority list. Finally, the graphics suite has also been updated to use Python 3 and Cartopy (v2018 used Python 2 and Basemap), which produces more user-friendly graphical products.

The track MAE is improved in v2021 compared to v2018 across the full sample for all lead times from 18–120 h by 3–5% for the unperturbed control member and by 2–8% for the ensemble mean. Track improvements in the West Pacific are even larger, with a reduction in error in the 5–10% range for the unperturbed control member. The northeast track bias in the West Pacific is also reduced in this update. Intensity MAE is improved by up to 9% in the ensemble control member and 10% in the ensemble mean at middle lead times (18–54 h). The intensity bias is also substantially reduced, with the weak bias at 36 h in v2018 reduced by greater than 50% in v2021. RI statistics indicate that v2021 has a better probability of detection, success rate, and threat score than v2018 from 0–90 h, and also predicts RI at a relative frequency closer to that observed in nature. The intensity distribution is also improved in v2021, with a greater ability to produce TCs at the high end of the distribution in the 135–160 kt range, as compared to v2018. Additionally, 34, 50, and 64 kt wind radii (R34, R50, R64) MAE and bias are all substantially improved in v2021. As an example of the substantial magnitude of the impact, the positive bias in 34 kt wind radii at 120 h is reduced from 25 n mi in v2018 to 9 n mi in v2021, corresponding to a 64% reduction in bias. The radius of maximum wind (Rmax) is also slightly improved in v2021 compared to v2018.

Several probabilistic verification metrics were also used to compare the COAMPS-TC ensemble v2021 to v2018, including the spread-skill relationship, uncertainty discrimination, and rank histograms. The spread-skill relationship for track in v2018 was already extremely well calibrated, and the spread-skill relationship for v2021 represents a slight degradation. This tradeoff, corresponding to a reduction in track spread greater than the reduction in mean track error, is at the cost of minimizing unrealistic outlier cases. By adjusting the perturbation initial condition and boundary condition coefficients, unrealistic outlier tracks have been substantially reduced. We also expect the spread-skill score for track of v2021 to improve when a future change to the grid 1 configuration, which has already been implemented in the v2021 deterministic COAMPS-TC, is implemented in the ensemble. Overall, the spread-skill relationship for intensity has not changed substantially in v2021 and is approximately equal to that of v2018.

Uncertainty discrimination analysis shows that there is a slight degradation for track, from 0.40 to 0.46, corresponding to a 15% increase. At early lead times of 24–48 h, performance in track uncertainty discrimination is better in v2018, while it is better at later lead times from 72–120 h in v2021. Uncertainty discrimination for intensity is improved in v2021 compared to v2018 at all lead times from 24–96 h, with a slight degradation at 120 h. Overall, the uncertainty discrimination score is improved from 0.53 to 0.41, representing a 23% reduction. Finally, rank histograms for intensity indicate that v2021 is more likely than v2018 to have all members with forecast intensity higher than the observed intensity. This tendency is most pronounced at early lead times. It is shown that v2021 has somewhat more relative frequency in the outermost bins, corresponding to an observation outside the ensemble envelope; however, the rank histogram in v2021 is more symmetric, indicating better centering of the distribution.

Finally, we will also seek feedback from JTWC regarding the new and improved graphics suite. It is expected that the improved figure clarity and aesthetics will be beneficial to forecasters issuing timely forecasts. Given the demonstrated improvement in skill, we expect the v2021 update to the COAMPS-TC ensemble to be well received by FNMOC and JTWC.

6. REFERENCES

1. P. Bougeault, “The Diurnal Cycle of the Marine Stratocumulus Layer: A Higher-order Model Study,” *J. Atmos. Sci.*, **42**, 2826–2843, 1985.

2. J. D. Doyle, J. Moskaitis, Y. Jin, W. Komaromi, et. Al, “Recent Progress and Challenges in Tropical Cyclone Intensity Prediction using COAMPS-TC,” *100th Annual Meeting of the AMS*, Boston, MA, 2020, Amer. Meteor. Soc., https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/363334.

3. J. D. Doyle, Y. Jin, R. Hodur, S. Chen, et. al, “Tropical Cyclone Prediction using COAMPS-TC,” *Oceanography*, **27**, 92–103, 2014.

4. J. D. Doyle, Y. Jin, R. Hodur, S. Chen, et. al, "Real-time Tropical Cyclone Prediction using COAMPS-TC," *Advances in Geosciences*, **28**, Eds. Chun-Chieh Wu and Jianping Gan, World Scientific Publishing Company, Singapore, 15-28, 2012.
5. Q. Fu and K.-N. Liou, "Parameterization of the Radiative Properties of Cirrus Clouds," *J. Atmos. Sci.*, **50**, 2008–2025, 1993.
6. Y. Jin, W.T. Thompson, S. Wang, and C.-S. Liou, "A Numerical Study of the Effect of Dissipative Heating on Tropical Cyclone Intensity," *Wea. Forecasting*, **22**, 950–966, 2007.
7. J. S. Kain and J. M. Fritsch, "Convective Parameterization for Mesoscale Models: The Kain-Fritsch Scheme. The Representation of Cumulus Convection in Numerical Models," *Meteor. Monogr.*, **46**, Amer. Meteor. Soc., 165–170, 1993.
8. W. A. Komaromi, P.A. Reinecke, J.D. Doyle, and J.R. Moskaitis, "The Naval Research Laboratory's Coupled Ocean–Atmosphere Mesoscale Prediction System–Tropical Cyclone Ensemble (COAMPS-TC Ensemble)," *Wea. Forecasting*, **36**, 499–517, 2021.
9. J. F. Louis, "A Parametric Model of Vertical Eddy Fluxes in the Atmosphere. *Bound.-Layer Meteor.*, **17**, 187–202, 1979.
10. J. Masters, J., "The Most Reliable Hurricane Models, According to their 2019 Performance. Yale Climate Connections," 2020, accessed 3 September 2020, https://yaleclimateconnections.org/2020/08/the-most-reliable-hurricane-models-according-to-their-2019-performance/.
11. G. Mellor and T. Yamada, "A Hierarchy of Turbulence Closure Models for Planetary Boundary Layers," *J. Atmos. Sci.*, **32**, 1278–1282, 1983.
12. S. A. Rutledge and P.V. Hobbs, "The Mesoscale and Microscale Structure of Organization of Clouds and Precipitation in Midlatitude Cyclones. VIII: A Model for the "Seeder-feeder" Process in Warm-frontal Rainbands," *J. Atmos. Sci.*, **40**, 1185–1206, 1983.
13. M. Tiedtke, "A Comprehensive Mass Flux Scheme for Cumulus Parametrization in Large-scale Models," *Mon. Wea. Rev.*, **117**, 1779–1800, 1989.
14. S. Wang, Q. Wang, and J. Doyle, "Some Improvement of Louis Surface Flux Parameterization. Preprints," 15th Symp. on Boundary Layers and Turbulence, Wageningen, Netherlands, Amer. Meteor. Soc., 547–550, 2002.