

Seasonal forecasts of North Atlantic tropical cyclone activity in the North American Multi-Model Ensemble

Article

Accepted Version

Manganello, J. V., Cash, B. A., Hodges, K. I. and Kinter III, J. L. (2017) Seasonal forecasts of North Atlantic tropical cyclone activity in the North American Multi-Model Ensemble. *Climate Dynamics*. ISSN 0930-7575 doi:
<https://doi.org/10.1007/s00382-017-3670-5> Available at
<http://centaur.reading.ac.uk/70252/>

It is advisable to refer to the publisher's version if you intend to cite from the work.

Published version at: <http://link.springer.com/article/10.1007/s00382-017-3670-5>

To link to this article DOI: <http://dx.doi.org/10.1007/s00382-017-3670-5>

Publisher: Springer

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the [End User Agreement](#).

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

1 **Seasonal Forecasts of North Atlantic Tropical Cyclone**
2 **Activity in the North American Multi-Model Ensemble**

3

4 Julia V. Manganello¹, Benjamin A. Cash¹, Kevin I. Hodges², James L. Kinter III^{1,3}

5

6 ¹ Center for Ocean-Land-Atmosphere Studies (COLA), George Mason University
7 (GMU), Fairfax, Virginia, USA

8 ² Dept. of Meteorology, University of Reading, Reading, UK

9 ³ George Mason University (GMU), Fairfax, Virginia, USA

10

11 For Submission to the NMME Special Issue in Climate Dynamics

12

13 Corresponding Author:
14 Julia V. Manganello
15 Center for Ocean-Land-Atmosphere Studies
16 113 Research Hall, Mail Stop 2B3
17 George Mason University
18 4400 University Drive
19 Fairfax, VA 22030 USA
20 jvisneva@gmu.edu
21 Phone: 703-993-5716
22 Fax: 703-993-5700

23

24 **Abstract**

25 The North American Multi-Model Ensemble (NMME)-Phase II models are evaluated
26 in terms of their retrospective seasonal forecast skill of the North Atlantic (NA)
27 tropical cyclone (TC) activity, with a focus on TC frequency. The TC identification
28 and tracking algorithm is modified to accommodate model data at daily resolution.
29 It is also applied to three reanalysis products at the spatial and temporal resolution
30 of the NMME-Phase II ensemble to allow for a more objective estimation of forecast
31 skill. When used with the reanalysis data, the TC tracking generates realistic
32 climatological distributions of the NA TC formation and tracks, and represents the
33 interannual variability of the NA TC frequency quite well.

34

35 Forecasts with the multi-model ensemble (MME) when initialized in April and later
36 tend to have skill in predicting the NA seasonal TC counts and TC days. At longer
37 leads, the skill is low or marginal, although one of the models produces skillful
38 forecasts when initialized as early as January and February. At short lead times,
39 while demonstrating the highest skill levels the MME also tends to significantly
40 outperform the individual models and attain skill comparable to the reanalysis. In
41 addition, the short-lead MME forecasts are quite reliable. It is found that the overall
42 MME forecast skill is limited by poor representation of the low-frequency variability
43 in the predicted NA TC frequency, and large fluctuations in skill on decadal time
44 scales. Addressing these deficiencies is thought to increase the value of the NMME
45 ensemble in providing operational guidance.

46 **1. Introduction**

47 Recognizing high socioeconomic significance of tropical cyclone (TC) prediction,
48 dynamical seasonal forecasts of TC activity have been pursued since the early 2000s
49 using low-resolution climate models (see reviews by Camargo et al. 2007; Camargo
50 and Wing 2016). These efforts have been gaining ground in recent years with the
51 improvements in the prediction systems including the increase of horizontal and
52 vertical resolutions of the component models (Molteni et al. 2011; Vecchi et al.
53 2014; Camp et al. 2015; Manganello et al. 2016) and wider use of ensemble
54 forecasting and multi-model ensemble approach (MME; Vitart 2006; Vitart et al.
55 2007). One such system is the North American Multi-Model Ensemble (NMME)
56 experimental multiagency seasonal forecasting system (Kirtman et al. 2014), which
57 is currently delivering real-time seasonal-to-interannual predictions used for
58 operational guidance. In the second stage of this project (NMME-Phase II),
59 improvements to the modeling and data assimilation systems have been introduced,
60 the size of forecast ensembles has increased, and more complete and higher
61 temporal frequency data has become available. In light of these developments, it
62 has become possible to evaluate the skill of dynamical seasonal forecasts of TC
63 activity by the individual NMME models and the corresponding MME to determine
64 whether these forecasts are skillful enough to be used in operational hurricane
65 outlooks.

66 In this paper, we examine the performance of the NMME-Phase II retrospective
67 forecasts of the North Atlantic (NA) seasonal mean TC activity where predicted
68 storms are identified directly in the model data using a feature-tracking algorithm.

69 Due to data limitations and relatively coarse horizontal resolution of the NMME
70 models (see Sections 2a and b), our analysis is largely limited to TC frequency, and
71 we briefly examine TC days¹ and regional TC activity as represented by track
72 density (see Vecchi et al. 2014; Manganello et al. 2016). For verification purposes,
73 we use three different reanalysis products in addition to the postseason best track
74 data, such as IBTrACS (see Section 2c). This is done to isolate the influence of model
75 resolution and the TC identification approach on the verification results. In addition
76 to assessing the overall level of skill, our goal is to identify aspects of the simulations
77 that could lead to potential improvements in the TC forecast skill and translate into
78 further developments of the NMME models.

79 Section 2 presents the NMME-Phase II models and hindcast datasets, and
80 introduces the observational and reanalysis data used to assess the skill of TC
81 hindcasts. It also describes the methodology of identifying and tracking the TCs in
82 the model data and reanalysis. Assessment of the seasonal forecast skill of the NA
83 TC activity, its dependence on the month of initialization and low-frequency
84 variability are presented in Section 3, along with a brief description of the
85 climatology of TC formation and tracks. Discussion of the results and concluding
86 remarks are included in Section 4.

87

¹ “TC days” is defined as a lifetime of all TCs accumulated over a season, measured in days.

88 **2. Data and Methods**

89 *a. NMME-Phase II models and data*

90 The NMME-Phase II ensemble consists of coupled prediction systems from North
91 American modeling centers and the Canadian Meteorological Centre (CMC). Table 1
92 contains information about the NMME-Phase II models and hindcast datasets used
93 in this study². The NMME System Phase II hindcast data is available for download
94 from the Earth System Grid at the National Center for Atmospheric Research (NCAR)
95 (<https://www.earthsystemgrid.org/search.html?Project=NMME>).

96 Atmospheric horizontal resolution of the models in Table 1 is relatively coarse
97 (between about 1 and 2 degrees), which is common to most present-day operational
98 seasonal prediction systems. (The output resolution is 1°x1° grid for all models.)
99 Daily frequency is the highest temporal output resolution for the majority of the
100 NMME-Phase II models. This rather coarse horizontal and temporal resolution of
101 the data puts additional constraints on the choices of objective criteria used for TC
102 identification, which is further elaborated below. A roughly 30-year period is
103 considered long enough to evaluate the skill of long-range predictions. The hindcast
104 start times include all 12 calendar months, which in addition to a large number of
105 lead times allows for an assessment of long-lead (forecasts initialized as early as
106 January) and short-lead (initialization as late as August) predictions.

107

108 *b. Tracking of tropical cyclones*

² At the time of this writing, daily dynamical fields for a common 1982-2012 hindcast period were available for download only for a subset of the NMME-Phase II models, which are listed in Table 1.

109 Identification and tracking of TCs in coarse- (horizontal) resolution models has
110 been done since the early 1980s, and a variety of methods exist to minimize the
111 effect of resolution on detection criteria (e.g., Walsh et al. 2007; Strachan et al.
112 2013). On the other hand, to resolve the TC trajectory, including its pre- and post-
113 TC stages, a sufficiently high temporal resolution is generally required with the 6-
114 hourly output frequency preferred for direct comparison with the best track data.
115 Tracking with daily data is not usually done, except in Smith et al. (2010) where TCs
116 are identified as minima in daily sea level pressure as they are tracked, which
117 reduces the number of possible matches but only captures the most intense part of
118 the lifecycle. In their study, the analysis is also restricted to the region between 0°
119 and 25°N. Recently, Vitart (2016) has successfully adjusted the tracking scheme
120 used at the European Centre for Medium-Range Weather Forecasts (ECMWF) to
121 evaluate the skill of sub-seasonal TC predictions using daily data.

122 In this study, the initial TC identification and tracking is based on the objective
123 feature-tracking methodology of Hodges (1995, 1999) and is tuned to work with
124 daily data, as opposed to 6-hourly data. The detection algorithm identifies vortices
125 as maxima in the 850-hPa relative vorticity field (in the Northern Hemisphere)
126 spectrally truncated at T42 with an intensity threshold of $1 \times 10^{-5} \text{ s}^{-1}$ and lifetimes
127 greater than 2 days (2 time steps). This tracking method allows TC tracks to be
128 captured in the deep tropics quite well but may underrepresent the extra-tropical
129 extensions of the tracks (see also Section 3a).

130 To separate predicted TCs from other synoptic-scale features, a set of TC
131 identification criteria needs to be applied to the raw tracks generated above. This

132 should include (1) a structural requirement of a warm core, (2) an intensity
133 threshold, along with (3) the formation region and (4) duration requirements. Due
134 to the coarseness of the spatial and temporal resolutions of the NMME-Phase II
135 models and limited availability of the surface wind data, we decided to base our TC
136 identification criteria solely on multi-level relative vorticity (at 850-hPa, 500-hPa
137 and 200-hPa levels common to all models in Table 1). To derive detection
138 thresholds in this case, simulated TC counts need to be calibrated against
139 observations. In this respect, our approach is similar to the method of Strachan et
140 al. (2013).

141 We have tested seven sets of TC identification criteria using May-November³
142 (MJJASON) reanalyses and model data (forecasts initialized in April). We varied the
143 number of levels used to define the vertical structure, assessed the sensitivity to the
144 presence of vorticity center at each level and monotonic reduction of vorticity with
145 height, and varied the minimum number of days when structural conditions need to
146 be satisfied (see Supplementary Material for more detail). In all cases, a warm core
147 condition remained the same, cyclogenesis was restricted to 0°-20°N over land and
148 0°-30°N over oceans, and 850-hPa vorticity at output resolution was used to
149 calibrate seasonal TC counts. For each reanalysis and NMME model, we have chosen
150 a set of TC identification criteria that maximizes their MJJASON TC frequency
151 correlation skill. These criteria are therefore not the same for all the datasets,
152 although the sensitivities are not large and are further discussed in the
153 Supplementary Material. While this is not a general practice, we believe that the

³ The MJJASON period encompasses most of the TC season in the NA basin.

154 above approach allows to better gauge the skill of each individual reanalysis and
155 model. These dataset-specific criteria do not change for the rest of the analysis,
156 including the skill assessment of long- and short-range predictions.

157

158 *c. Observational and reanalysis data*

159 For comparison with observations, we use data from the International Best
160 Track Archive for Climate Stewardship (IBTrACS, version v03r07; Knapp et al. 2010;
161 available online at <https://www.ncdc.noaa.gov/ibtracs/>). IBTrACS makes available
162 for public use a global dataset of post season analysis of TC position and intensity
163 (also know as “best track”) by merging storm information from multiple centers into
164 one product. The observed tracks are further processed here by retaining systems
165 with lifetimes greater than 2 days, of tropical storm strength for at least 1 day and
166 with first identification occurring between 0°-20°N over land and 0°-30°N over
167 oceans, to be more in line with the model and reanalysis tracks (see Section 2b). We
168 also use sea surface temperature (SST) data from the National Oceanic and
169 Atmospheric Administration (NOAA) Optimum Interpolation SST version 2 data set
170 (OISSTv2; Reynolds et al. 2002).

171 Since our choice of TC identification criteria (Section 2b) does not imply a close
172 match with the observational ones, it is prudent to use reanalysis data for more
173 direct verification of model results. In reanalyses, historical observations are
174 objectively ingested into the models with a goal to produce a consistent estimate of
175 the state of the climate. As such, reanalyses have an advantage of models by
176 providing a more comprehensive dataset. They are constrained by the observations

177 but limited by the raw input data and its quality, the resolution of the models used,
178 and the capabilities of the data assimilation system. Overall, applying the same
179 tracking methodology to the reanalysis and model data of the same spatial and
180 temporal resolution would allow a more objective estimation of the model skill.

181 We have used the following three reanalysis datasets: the National Centers for
182 Environmental Prediction (NCEP) Climate Forecast System Reanalysis (CFSR; Saha
183 et al. 2010); the Interim ECMWF Re-Analysis (ERA-I; Dee et al. 2011); and the
184 National Aeronautics and Space Administration (NASA) Modern Era Retrospective-
185 Analysis for Research and Applications (MERRA; Rienecker et al. 2011). The spatial
186 resolution of all reanalysis data was downgraded to the $1^{\circ}\times 1^{\circ}$ grid of the NMME-
187 Phase II model data. The temporal resolution was converted to daily, and the period
188 of 1982-2014 was used for analysis.

189

190 **3. Results**

191 *a. Climatologies of TC formation and tracks*

192 Prior to evaluating the skill of TC frequency forecasts, we verify whether the TC
193 identification and tracking approach chosen here generates realistic distributions of
194 genesis locations and tracks. Figs. 1 and 2 show NA genesis and track densities,
195 respectively, for the IBTrACS, reanalyses and the NMME-Phase II retrospective
196 seasonal forecasts. Reanalysis products reproduce main features of the genesis
197 pattern quite well, with varying levels of success depending on the specific
198 cyclogenesis center (Figs. 1a-d). CFSR is most accurate in representing the Main
199 Development Region (MDR; 10° - 25° N, 80° - 20° W), whereas in ERA-I and MERRA,

200 activity in this area is largely concentrated near the west coast of Africa. (Origin of
201 some tracks over West Africa is likely related to their tropical easterly wave
202 precursors being captured by the tracking algorithm (see also Manganello et al.
203 2012). For the same reason, the bulk of the MDR genesis is shifted further to the
204 east compared to observations.) The Gulf of Mexico center is underrepresented in
205 all reanalysis products, whereas the western Atlantic center is quite realistic across
206 the board. The Caribbean genesis is shifted southeast and is somewhat overactive in
207 ERA-I. This shift has been noted earlier and linked to the coarse spatial resolution of
208 the models (Manganello et al. 2012, 2016). The associated track density is overall
209 well reproduced (Figs. 2a-d), except in the extra-tropics which is likely a
210 consequence of tracking using daily data (see Section 2b).

211 Predicted genesis and track densities on the whole are less realistic compared to
212 observations and reanalyses, where formation regions are strongly concentrated in
213 space (Figs. 1e-h), and track density is overpredicted and too zonal in the tropics
214 and quite weak further north (Figs. 2e-h). However, the MDR genesis is rather
215 active in all the hindcasts, and other centers are well defined, except for the Gulf of
216 Mexico and the western Atlantic centers being absent in the CanCM3 forecasts. In
217 addition, the Gulf of Mexico center, where present, is more realistic than in the
218 reanalysis. On the other hand, the Caribbean genesis is too strong, and the
219 associated tracks are largely confined to the northern tip of South America. To
220 summarize, the tracking algorithm is capable of generating climatologies of the NA
221 TC formation and tracks with many realistic features, particularly when applied to
222 reanalysis products.

223

224 *b. April forecasts of the North Atlantic seasonal mean TC activity*

225 1). TC frequency

226 Fig. 3 shows the interannual variability of the observed and reanalyses-based NA
227 TC frequency, which is another demonstration of the utility of the TC tracking
228 method in estimating seasonal mean TC activity using daily data. The reanalysis
229 datasets reproduce the interannual variability quite well, with major peaks of 1995
230 and 2005 to the most part realistically represented. The correlation coefficients
231 between the reanalyses and the observed time series are also quite high ranging
232 from 0.67 to 0.81 (see Table 2). The reanalyses do differ considerably in terms of
233 their skill in representing multidecadal changes characterized by low activity in the
234 1980s and early 1990s and high activity in the latter part of the record (e.g.,
235 Goldenberg et al. 2001). ERA-I is the most successful in capturing this trend,
236 whereas CFSR displays no trend (see Fig. 3).

237 Retrospective correlation skill varies markedly among the NMME-Phase II
238 models (see Table 2 for MJJASON forecasts initialized in April). It is quite high for
239 CCSM4 and CanCM4 and is in fact similar to the skill of experimental high-
240 atmospheric-resolution coupled prediction systems in Project *Minerva* (Manganello
241 et al. 2016), whereas it is close to zero for GEOS-5 and CanCM3. As a consequence,
242 correlation of the MME mean⁴ is significant but rather modest and does not exceed

⁴ The MME mean is defined as the average over all the hindcasts, with all ensemble members of each model having equal weight.

243 the skill of all models in the ensemble. The root-mean-square error⁵ (RMSE), which
244 a measure of forecast accuracy, is fairly large, although the differences are not major
245 when the MME mean is compared to reanalyses (Table 3). RMSE for the detrended
246 time series is smaller across the board suggesting that low-frequency variability is
247 not well reproduced in the forecasts (see below). For short-range predictions, the
248 overall skill improves, and the advantages of the MME approach become more
249 evident (see Secion 3d).

250 A natural question arises whether the individual NMME-Phase II models are
251 indeed more or less skillful than their MME mean, and whether these models
252 including the MME display skill that is significantly different from the skill based on
253 the reanalyses data. The correlation coefficient is not considered a very good
254 measure to compare skill, as the presence of noise may lead to large differences in
255 this quantity. It is found that the squared error is a more appropriate metric
256 (DelSole and Tippett, 2014), and we choose the Wilcoxon signed-rank test for the
257 forecast skill comparison since it is not sensitive to the type of distribution (ibid.).
258 We find that at the 95% confidence level, the differences in skill among the four
259 NMME models and their MME mean are insignificant, except that the skill of GEOS-5
260 and CanCM3 is significantly lower that the skill of CanCM4. We also find that all
261 NMME models and the MME mean are as skillful as CFSR and ERA-I but less skillful
262 than MERRA. (The skill of CanCM3 is also significantly lower compared to ERA-I). It

⁵ Forecasts are calibrated (without cross-validation) where each ensemble member is multiplied by a constant factor so that the predicted ensemble-mean and observed climatologies become equal.

263 is worth emphasizing that the above skill comparison is based on the MJJASON
264 season (forecasts initialized in April).

265 Ensemble forecasts have an additional advantage of being able to quantify
266 uncertainty based on the probabilistic approach. One such measure is statistical
267 reliability, which can be expressed as a ratio of the ensemble spread and the RMSE
268 (SPRvERR). In a perfectly reliable ensemble forecast, forecast probabilities match
269 the observed frequencies, and the SPRvERR is equal to one. Individual NMME and
270 the MME mean April forecasts are found to be underdispersed (or overconfident;
271 Table 4). Detrending the time series enhances reliability quite a bit which indicates
272 that poor low-frequency variability of the predicted NA TC frequency is indeed a
273 distinct source of forecast error. These results are similar to our findings in Project
274 *Minerva* (Manganello et al. 2016).

275 To further illustrate the above results, Fig.4 shows seasonal mean TC frequency
276 predicted by the CCSM4 and CanCM4 models along with their ensemble information
277 compared with observations. Both models capture year-to-year fluctuations quite
278 well, particularly in the 1990s and early 2000s where only several seasons fall
279 outside the 10th-90th percentile range (1992, 1997, and 2005 for CCSM4; and 1992,
280 1995, 1997 and 2005 for CanCM4). Neither of the models reproduces the secular
281 trend, and the hindcast skill appears to be inferior in the 1980s and 2010s, which is
282 further discussed below.

283 2) TC days and TC track density

284 Seasonally accumulated lifetime of all TCs in the basin, or “TC days” (see
285 definition in Section 1), exhibits retrospective correlation skill behavior quite

286 comparable to TC frequency (Table 5). The forecasts that are skillful in predicting
287 TC frequency are to the most part also skillful in predicting TC days. For MJJASON
288 forecasts initialized in April the correlation of the MME mean TC days is not high but
289 significant (0.46), and increases to 0.59 at shorter leads (July and August
290 initializations). It is curious that reanalyses reproduce variability of TC days
291 seemingly better than TC frequency (using current tracking), where correlation for
292 TC days doesn't drop below 0.76 (Table 5).

293 One of the current challenges of seasonal TC forecasting is to provide regional
294 information, such as local TC occurrence or probability of landfall, which is more
295 relevant for decision-making (e.g., Vecchi et al. 2014; Camp et al. 2015; Manganello
296 et al. 2016; Murakami et al. 2016). Here we examine whether MME forecasts of the
297 NA TC activity have retrospective skill on sub-basin scales using track density as a
298 metric and Spearman rank correlation as a measure of performance (see
299 Manganello et al. 2016 for more detail). We compare this skill to the rank
300 correlation between the seasonal mean observed and reanalyses-derived track
301 densities. All three reanalysis products are quite successful at reproducing
302 interannual variability of regional TC activity over most of the NA domain (Figs. 5a-
303 c). The regions with significant correlations common to all products are the MDR,
304 the Caribbean Sea, the Gulf of Mexico and central subtropical North Atlantic. These
305 regions also tend to show the highest correlation values. The results do not seem to
306 be particularly sensitive to whether the extended MJJASON season or the peak ASON
307 season is examined (Figs. 5e-g). In comparison, for the longer-lead MME forecasts
308 initialized in April the regions with significant skill are rather sparse and limited to

309 some parts of the MDR and the westernmost margins of the Caribbean Sea and the
310 Gulf of Mexico (Fig. 5d). The absence of any skill north of about 30°N is likely
311 related to strong underprediction of climatological tracks at these latitudes in the
312 NMME models (see Section 3a). At shorter leads (MME forecasts initialized in July),
313 the region with significant skill markedly increases and now covers the western part
314 of the MDR and the whole Caribbean Sea (Fig. 5h). Fairly high retrospective forecast
315 skill in the vicinity of Caribbean islands suggests that predictions of TC landfall
316 frequency in this region may also be skillful. Overall, the skill of regional TC activity
317 forecasts in the NMME is rather modest compared to other coupled prediction
318 systems that employ atmospheric models with much higher horizontal resolution
319 (see Vecchi et al. 2014; Manganello et al. 2016; Murakami et al. 2016).

320

321 *c. Low-frequency variability in prediction skill*

322 The NMME-Phase II ensemble exhibits variability in the retrospective forecast
323 skill of the NA TC frequency (Fig. 6). Compared to the reanalyses, which maintain
324 relatively constant skill throughout the hindcast period, the MME mean displays
325 markedly lower skill in the 1980s and early 1990s, and also late 2000s and 2010s
326 (Fig. 6a). During these two periods, the model skill deviates from the reanalyses. In
327 contrast, it is quite comparable to the reanalyses in the late 1990s and early 2000s.
328 Since the NA TC season peaks in August-October, forecasts initialized in June could
329 be considered short-lead forecasts of the full hurricane season. We find that at
330 shorter leads (Fig. 6b), forecast skill becomes more in line with the reanalyses in the

331 latter part of the record. This tendency is also present in forecasts initialized in May
332 (not shown).

333 Loss of skill in the 1980s is not unique to the NMME-Phase II models. Similar
334 behavior was also found in all *Minerva* hindcasts (Manganello et al. 2016) where it
335 was linked to more deficient initialization of ocean fields. It is also feasible that
336 predictability of the NA TC activity can fluctuate from one decade to another. The
337 influence of certain climatic factors that serve as predictors of the NA TC activity
338 may depend on the underlying climate conditions (Fink et al. 2010; Caron et al.
339 2015). Current seasonal prediction systems are perhaps able to reproduce some of
340 the relationships but not others or do not time them correctly, which may
341 contribute to the drop in skill.

342 While a detailed analysis of these influences is beyond the scope of the current
343 paper, as a first step we examine here the relationship between the NMME forecasts
344 of TC frequency and several well established predictors of the NA TC genesis, and
345 compare results to observations and reanalyses. The selected climate indices are: 1)
346 SST averaged over the MDR; 2) relative SST index⁶, and 3) the Niño-3.4 index⁷ (see,
347 e.g., Villarini et al. 2010; Vecchi et al. 2011; Caron et al. 2015 and the extensive lists
348 of references in these papers). Both observations and reanalyses suggest a stronger
349 relationship between the MDR SSTs and the NA TC frequency in the late 1990s and
350 early 2000s compared to the earlier and latter parts of the record where
351 correlations become marginally significant (Fig. 7a). The correlation with the

⁶ Relative SST index is defined as the difference between MDR SST and global tropical-mean SST (e.g., Zhao et al. 2010).

⁷ Niño-3.4 index is defined as SST averaged over 5°S-5°N, 120°-170°W.

352 relative SST index is higher and more constant throughout the time period (Fig. 7b),
353 as is the negative connection with the El Niño and the Southern Oscillation (ENSO)
354 except perhaps in 2000s where reanalyses data suggest a weakening of this
355 relationship (Fig. 7c). The NMME models and their MME mean tend to display
356 rather different behavior. During the earlier and latter parts of the hindcast period,
357 TC frequency forecasts appear to be much stronger driven by variations in the
358 predicted MDR SSTs and the relative SST index compared to the middle part of the
359 record, opposite to what observations and reanalyses demonstrate (Figs. 7a and b).
360 It is curious that the late 1990s and early 2000s when the MME correlations with
361 the MDR SSTs and the relative SST index are most realistic coincide with the period
362 of the highest MME TC frequency forecast skill (Fig. 6a). On the other hand, the rest
363 of the hindcast period when these correlations are too high and markedly outside
364 the range of the observed/reanalyses values is also when the forecast skill is at the
365 lowest levels as described above and shown in Fig. 6a. In addition, the retrospective
366 forecast skill of the MDR and relative SST indices is generally quite high except in
367 the 1980s and early 1990s when forecasts of the relative SST index are not skillful
368 (see Fig. S1 in the Supplementary Material). This could further limit the quality of
369 the TC frequency predictions during this time period. In contrast, the influence of
370 ENSO appears to be captured quite well by the MME forecasts, except possibly in the
371 1980s and late 2000s when it appears to be somewhat stronger (Fig. 7c); the
372 hindcast skill of the Niño-3.4 index is the highest among the indices examined and
373 also fairly constant throughout the record (Fig. S1).
374

375 *d. Long- and short-lead forecasts*

376 The NA TC hindcast skill as a function of the initialization month is shown in Fig.
377 8, along with the results for the reanalyses and measures of “null skill”. At longer
378 lead times (earlier than April), the MME mean shows marginal skill when initialized
379 in February relative to the IBTrACS trailing 5-yr average, which is a skill metric
380 recommended by the World Meteorological Organization (WMO 2008; Fig. 8a). In
381 this reference forecast, the interannual variability is smoothed out but the
382 interdecadal variability is preserved to some extent. The best performing forecasts
383 at long leads are produced by CanCM4 and are skillful for January and February
384 initializations. It is notable that for most models and the MME mean the skill curves
385 in Fig. 8a display substantial variability from month to month. This “noisiness” is
386 largely due to low-frequency variability being forecasted at varying levels of skill
387 depending on the initialization month. (Compare also with Fig. 8b that shows
388 similar metrics computed for the detrended time series and displaying a more
389 consistent increase in skill with lead time.) Relative to persistence, or the previous
390 season’s TC count, the detrended MME mean shows no long-lead skill except
391 perhaps when initialized in March. All detrended long-lead CanCM4 forecasts show
392 skill albeit marginal.

393 When the hurricane season is approached (March and June initializations) the
394 skill drops somewhat (Figs. 8a and b). At short lead times (July and August), it
395 rebounds and displays the highest levels overall (see also Table 2). It is notable that
396 all detrended MME mean forecasts initialized in April and later are consistently
397 skillful relative to persistence (Fig. 8b). The short-lead MME mean correlation skill

398 (RMSE) also shows the highest (lowest) value among all the models (detrended
399 only; see Tables 2 and 3). In addition, it becomes comparable to the skill of the
400 reanalyses. For instance, RMSEs of forecasts initialized in July are lower than for
401 CFSR and ERA-I (detrended only in the latter case; Table 3). The short-lead MME
402 mean forecasts are also quite reliable, although somewhat over-dispersed when
403 detrended (Table 4). It is curious that among the forecasts initialized in June
404 through August the best performing model is CanCM3, whereas it is one of the worst
405 performing at longer leads. If April forecasts were chosen as a benchmark and the
406 MME are based on two models with skill (CCSM4 and CanCM4), the resultant
407 correlation at short leads is markedly lower compared to the MME based on *all*
408 available models (not shown). This is one of the advantages of the multi-model
409 ensemble approach that is not always obvious.

410 The skill of the MME mean relative to the individual NMME-Phase II models and
411 the reanalyses is further assessed using the difference between the squared error as
412 a skill metric and testing the significance by applying the Wilcoxon signed-rank test
413 (see Fig. 9; DelSole and Tippett, 2014). In the vast majority of cases, the MME mean
414 outperforms the individual model with differences being statistically significant at
415 short lead times (June and July initializations). Relative to the reanalyses, the MME
416 mean shows larger error most of the time (except at short leads with respect to
417 CFSR), although it is significant primarily at long leads and when compared to
418 MERRA only. It is also notable that at most lead times, the reliability is improved
419 slightly for the MME mean and to a larger extent when the time series are detrended
420 (not shown).

421

422 **4. Summary and conclusions**

423 In this study, the NMME-Phase II models are interrogated in terms of the
424 retrospective seasonal forecast skill of the NA TC frequency. The TCs are identified
425 explicitly in the model data by means of an objective feature-tracking methodology.
426 Due to the synoptic nature of these storms, daily resolution (the highest available
427 for the ensemble) is generally considered coarse for TC tracking. As part of this
428 work, we have adjusted the TC identification and tracking algorithm to work with
429 daily data and also applied it to three reanalysis products (CFSR, ERA-I and MERRA)
430 that were coarsened to have the same spatial and temporal resolution of the NMME-
431 Phase II ensemble. The latter step provides additional verification data (apart from
432 best track data) where the effects of resolution and the TC identification approach
433 have been isolated which allows for a more objective estimation of forecast skill.

434 The TC tracking method used here, when applied to reanalysis data, produces
435 realistic climatological distributions of the NA TC formation and tracks. Low track
436 density in the extra-tropics is a common deficiency, which is a result of tracking
437 using daily data. The tracking is also quite skillful in reproducing the interannual
438 variability of the TC frequency relative to the IBTrACS with correlations ranging
439 between 0.67 and 0.81 depending on the reanalysis product. These values are quite
440 comparable to the estimates obtained in Strachan et al. (2013) and Roberts et al.
441 (2015) where both studies utilized six-hourly data.

442 Long-lead (March and earlier) retrospective seasonal forecasts of the NA TC
443 frequency with the MME based on the available NMME-Phase II models are found to

444 have low or marginal skill, although one of the models (CanCM4) produces skillful
445 forecasts when initialized as early as in January and February. At shorter leads
446 (April and later), the MME mean forecasts are largely skillful with the best
447 performance for July and August initializations. Skill metrics evaluated for the
448 detrended time series display a more systematic increase in skill with shorter lead
449 time, and all detrended MME mean forecasts initialized in April and later are
450 consistently skillful. At short lead times (June through August), the MME mean also
451 tends to significantly outperform the individual models and attain skill comparable
452 to the reanalysis. The short-lead MME mean forecasts are also quite reliable, while
453 being under-dispersed at longer leads.

454 We have identified several deficiencies in the simulations that likely limit the
455 NMME-Phase II seasonal hindcast skill of the NA TC frequency.

456 1. None of the models or the MME mean independent of the initialization month
457 can realistically represent low-frequency variability characterized by low
458 activity in the 1980s and early 1990s and higher activity thereafter. The skill
459 metrics computed for the detrended time series show higher scores in the
460 vast majority of cases. This suggests that poor multi-year variability in the
461 forecasts may indeed be a source of forecast error. This problem is not trivial
462 and is characteristic of other prediction systems like *Minerva* (Manganello et
463 al. 2016) and several reanalysis products, e.g., MERRA and CFSR. It could be
464 related, for instance, to poor skill in reproducing downward trends in upper
465 tropospheric temperature (Emanuel et al. 2013; Vecchi et al. 2013),
466 inadequate representation of the effects of aerosols and ozone (Evan et al.

467 2009, 2011; Emanuel et al. 2013), possibly deficiencies in simulating tropical
468 heating and atmospheric teleconnections (Manganello et al. 2016), and the
469 sensitivity to the identification of weak and short-lived TCs in the model and
470 reanalysis data.

471 2. We have shown that the MME mean forecasts exhibit a large drop in skill in
472 the 1980s and early 1990s and also late 2000s and 2010s (mostly at longer
473 leads). It is curious that during the rest of the period (late 1990s and early
474 2000s), the MME mean skill is quite comparable to the reanalyses, which
475 maintain relatively constant skill throughout the hindcast time period. Early
476 in the record, forecast errors could be partly related to deficiencies in the
477 model initialization. Although the problem as a whole may be more complex
478 and indicate that certain physical relationships that underline predictability
479 of the NA TC activity may not be consistently reproduced or properly timed.

480 Addressing the above issues, while not an easy task, could lead to marked
481 improvements in the seasonal forecast skill and increase the value of the NMME
482 ensemble in providing operational guidance.

483

484 *Acknowledgements.* Funding of COLA for this study is provided by grants from NSF
485 (AGS-1338427), NOAA (NA09OAR4310058 and NA14OAR4310160), NASA
486 (NNX14AM19G), and the ONR grant (N00014-15-1-2745). We acknowledge NOAA
487 MAPP, NSF, NASA, and the DOE that support the NMME-Phase II system, and we
488 thank the climate modeling groups (Environment Canada, NASA, NCAR,
489 NOAA/GFDL, NOAA/NCEP, and University of Miami) for producing and making

490 available their model output. NOAA/NCEP, NOAA/CTB, and NOAA/CPO jointly
491 provided coordinating support and led development of the NMME-Phase II system.
492 We also gratefully acknowledge computing resources on the Yellowstone
493 supercomputer provided by the National Center for Atmospheric Research.
494

495 **References**

- 496 Camargo, S. J., A. G. Barnston, P. J. Klotzbach, and C. W. Landsea, 2007: Seasonal
497 tropical cyclone forecasts. *World Meteorological Organization Bulletin*, **56**, 297-
498 309.
- 499 Camargo, S. J., and A. A. Wing, 2016: Tropical cyclones in climate models. *WIREs*
500 *Climate Change*, **7**, 211-237.
- 501 Camp, J., M. Roberts, C. MacLachlan, E. Wallace, L. Hermanson, A. Brookshaw, A.
502 Arribas, A. A. and Scaife, 2015: Seasonal forecasting of tropical storms using the
503 Met Office GloSea5 seasonal forecast system. *Q.J.R. Meteorol. Soc.*, doi:
504 10.1002/qj.2516.
- 505 Caron, L.-P., M. Boudreault, and C. L. Bruyere, 2015: Changes in large-scale controls
506 of Atlantic tropical cyclone activity with the phases of the Atlantic multidecadal
507 oscillation. *Clim. Dyn.*, **44**, 1801-1821.
- 508 Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: configuration and
509 performance of the data assimilation system. *Q.J.R. Meteorol. Soc.*, **137**, 553–597.
510 doi: 10.1002/qj.828
- 511 DelSole, T., and M. K. Tippett, 2014: Comparing Forecast Skill, *Mon. Weather Rev.*,
512 **142(12)**, 4658–4678.
- 513 Emanuel, K. A., S. Solomon, D. Folini, S. Davis, and C. Cagnazzo, 2013: Influence of
514 tropical tropopause layer cooling on Atlantic Hurricane activity. *J. Clim.*, **26**,
515 2288–2301.
- 516 Evan, A. T., G. R. Foltz, D. Zhang, and D. J. Vimont, 2011: Influence of African dust on
517 ocean-atmosphere variability in the tropical Atlantic. *Nature Geoscience*, **4**, 762-

518 765.

519 Evan, A. T., D. J. Vimont, A. K. Heidinger, J. P. Kossin, and R. Bennartz, 2009: The Role
520 of Aerosols in the Evolution of Tropical North Atlantic Ocean Temperature
521 Anomalies. *Science*, **324**, 778-781.

522 Fink, A. H., J. M. Schrage, and S. Kotthaus, 2010: On the potential causes of the
523 nonstationary correlations between West African Precipitation and Atlantic
524 Hurricane Activity. *J. Climate*, **23**, 5437-5456.

525 Goldenberg, S. B., C. W. Landsea, A. M. Mestas-Nu.ez, and W. M. Gray, 2001: The
526 recent increase in Atlantic hurricane activity: Causes and implications. *Science*,
527 **293**, 474-479.

528 Hodges, K. I., 1995: Feature tracking on the unit sphere. *Mon. Wea. Rev.*, **123**, 3458-
529 3465.

530 Hodges, K. I., 1999: Adaptive constraints for feature tracking. *Mon. Wea. Rev.*, **127**,
531 1362-1373.

532 Kirtman, B.P., and Coauthors, 2014: The North American multimodel ensemble:
533 phase 1 seasonal-to-interannual prediction; phase-2 toward developing
534 intraseasonal prediction. *Bull. Am. Meteorol. Soc.*, **95**, 585-601.

535 Knapp, K. R., M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann, 2010: The
536 International Best Track Archive for Climate Stewardship (IBTrACS). *Bull. Amer.*
537 *Meteor. Soc.*, **91**, 363-376.

538 Manganello, J. V., K. I. Hodges, B. A. Cash, J. L. Kinter III, E. L. Altshuler, M. J. Fennessy,
539 F. Vitart, F. Molteni, and P. Towers, 2016: Seasonal Forecasts of Tropical Cyclone

540 Activity in a High Atmospheric Resolution Coupled Prediction System, *J. Climate*,
541 **29**, 1179-1200.

542 Manganello, J. V., K. I. Hodges, J. L. Kinter III, B. A. Cash, L. Marx, T. Jung, D.
543 Achuthavarier, J. M. Adams, E. L. Altshuler, B. Huang, E. K. Jin, C. Stan, P.
544 Towers and N. Wedi, 2012: Tropical Cyclone Climatology in a 10-km Global
545 Atmospheric GCM: Toward Weather-Resolving Climate Modeling. *J. Climate*, **25**,
546 3867-3893.

547 Merryfield, W. J., and Coauthors, 2013: The Canadian seasonal to interannual
548 prediction system. Part I: Models and initialization. *Mon. Wea. Rev.*, **141**, 2910–
549 2945.

550 Molteni, F., T. Stockdale, M. Balmaseda, G. Balsamo, R. Buizza, L. Ferranti, L.
551 Magnusson, K. Mogensen, T. Palmer, and F. Vitart, 2011: The new ECMWF
552 seasonal forecast system (System 4). *ECMWF Technical Memorandum*, No. **656**,
553 pp. 49.

554 Murakami, H., G. A. Vecchi, G. Villarini, T. L. Delworth, R. Gudgel, S. Underwood, X.
555 Yang, W. Zhang, and S.-J. Lin, 2016: Seasonal forecasts of major hurricanes and
556 landfalling tropical cyclones using a high-resolution GFDL coupled climate
557 model. *J. Climate*, **29**, 7977-7989.

558 Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, 2002: An
559 improved in situ and satellite SST analysis for climate. *J. Climate*, **15**, 1609-1625.

560 Rienecker, M. M., and Coauthors, 2011: MERRA: NASA's Modern-Era Retrospective
561 Analysis for Research and Applications. *J. Climate*, **24**, 3624–3648.

562 Roberts, M. J., P. L. Vidale, M. S. Mizieliński, M.-E. Demory, R. Schiemann, J. Strachan,

563 K. Hodges, R. Bell, and J. Camp, 2015: Tropical Cyclones in the UPSCALE
564 Ensemble of High-Resolution Global Climate Models. *J. Climate*, **28**, 574-596.

565 Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull.*
566 *Am. Meteorol. Soc.*, **91(8)**, 1015-1057.

567 Smith, D. M., R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and A. A.
568 Scaife, 2010: Skilful multi-year predictions of Atlantic hurricane frequency.
569 *Nature Geoscience*, **3**, 846-849.

570 Strachan, J., Vidale, P. L., Hodges, K., Roberts, M., and Demory, M.-E., 2013:
571 Investigating global tropical cyclone activity with a hierarchy of AGCMs: the role
572 of model resolution. *J. Climate*, **26**, 133-152.

573 Vecchi, G. A., T. Delworth, R. Gudgel, S. Kapnick, A. Rosati, A. T. Wittenberg, F. Zeng,
574 W. Anderson, V. Balaji, K. Dixon, L. Jia, H.-S. Kim, L. Krishnamurthy, R. Msadek, W.
575 F. Stern, S. D. Underwood, G. Villarini, X. Yang, and S. Zhang, 2014: On the
576 Seasonal Forecasting of Regional Tropical Cyclone Activity. *J. Climate*, **27**, 7994-
577 8016.

578 Vecchi, G. A., S. Fueglistaler, I. M. Held, T. R. Knutson, and M. Zhao, 2013: Impacts of
579 atmospheric temperature trends on tropical cyclone activity. *J. Clim.*, **26**, 3877-
580 3891.

581 Vecchi, G. A., M. Zhao, H. Wang, G. Villarini, A. Rosati, A. Kumar, I. M. Held, and R.
582 Gudgel, 2011: Statistical-dynamical predictions of seasonal North Atlantic
583 hurricane activity. *Mon. Wea. Rev.*, **139**, 1070-1082.

584 Vernieres, G., C. Keppenne, M.M. Rienecker, J. Jacob, and R. Kovach, 2012: The GEOS-
585 ODAS, description and evaluation. NASA Technical Report Series on Global

586 Modeling and Data Assimilation, NASA/TM-2012-104606, Vol. 30.

587 Villarini, G., G. A. Vecchi, and J. A. Smith, 2010: Modeling the dependence of tropical
588 storm counts in the North Atlantic basin on climate indices. *Mon. Wea. Rev.*, **138**,
589 2681-2705.

590 Vitart, F., 2006: Seasonal forecasting of tropical storm frequency using a multi-
591 model ensemble. *Q. J. R. Meteorol. Soc.*, **132**, 647-666.

592 Vitart, F. 2016: Tropical cyclogenesis in the S2S Database. *S2S News Letter*, **No. 3**, 3-
593 6.

594 Vitart, F., M. R. Huddleston, M. Déqué, D. Peake, T. N. Palmer, T. N. Stockdale, M. K.
595 Davey, S. Ineson, and A. Weisheimer, 2007: Dynamically-based seasonal
596 forecasts of Atlantic tropical storm activity issued in June by EUROSIP. *Geophys.*
597 *Res. Lett.*, **34**, L16815, doi:10.1029/2007GL030740.

598 Walsh, K. J. E., M. Fiorino, C. W. Landsea, K. L. McInnes, 2007: Objectively
599 Determined Resolution-Dependent Threshold Criteria for the Detection of
600 Tropical Cyclones in Climate Models and Reanalyses. *J. Climate*, **20**, 2307-2314.

601 WMO, 2008: Report from expert meeting to evaluate skill of tropical cyclone
602 seasonal forecasts. World Meteorological Organization. Tech. Doc. 1455,
603 Geneva, Switzerland. 27 pp.

604 Zhao, M., I. M. Held, and G. A. Vecchi, 2010: Retrospective forecasts of the hurricane
605 season using a global atmospheric model assuming persistence of SST
606 anomalies. *Mon. Wea. Rev.*, **138**, 3858-3868.

607

608
 609
 610
 611
 612
 613
 614
 615
 616
 617
 618
 619
 620
 621
 622
 623
 624
 625
 626
 627
 628
 629
 630
 631
 632
 633
 634
 635
 636
 637
 638
 639
 640
 641
 642
 643
 644
 645
 646
 647
 648
 649
 650
 651
 652
 653

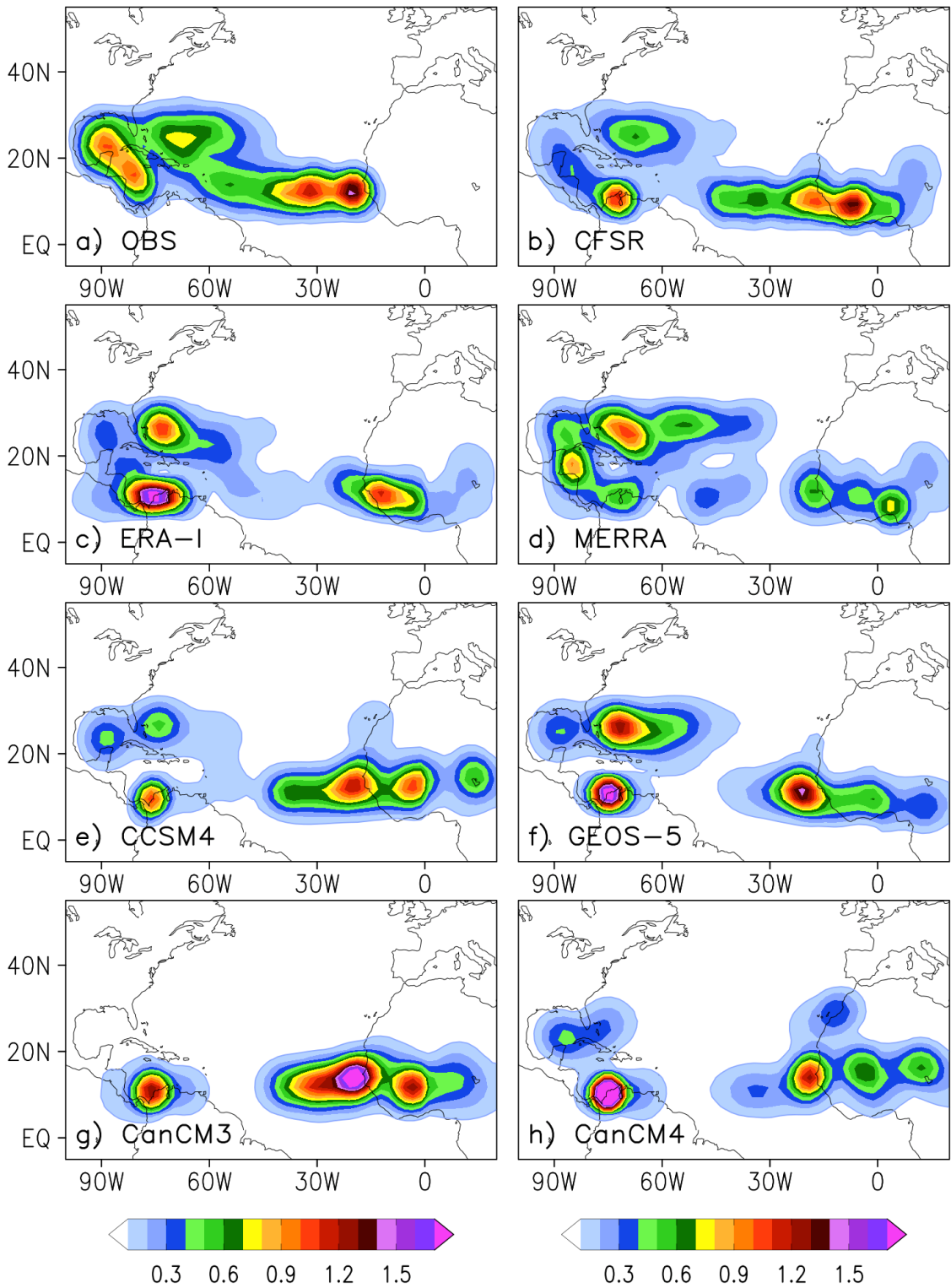


Figure 1: NA genesis densities for the MJJASON season as number density per season per unit area equivalent to a 5° spherical cap for (a) IBTrACS (OBS), (b) CFSR, (c) ERA-I, and (d) MERRA reanalyses based on 1982-2014, and (e) CCSM4, (f) GEOS-5, (g) CanCM3, and (h) CanCM4 seasonal hindcasts (all ensemble members) based on the time periods listed in Table 1.

654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698

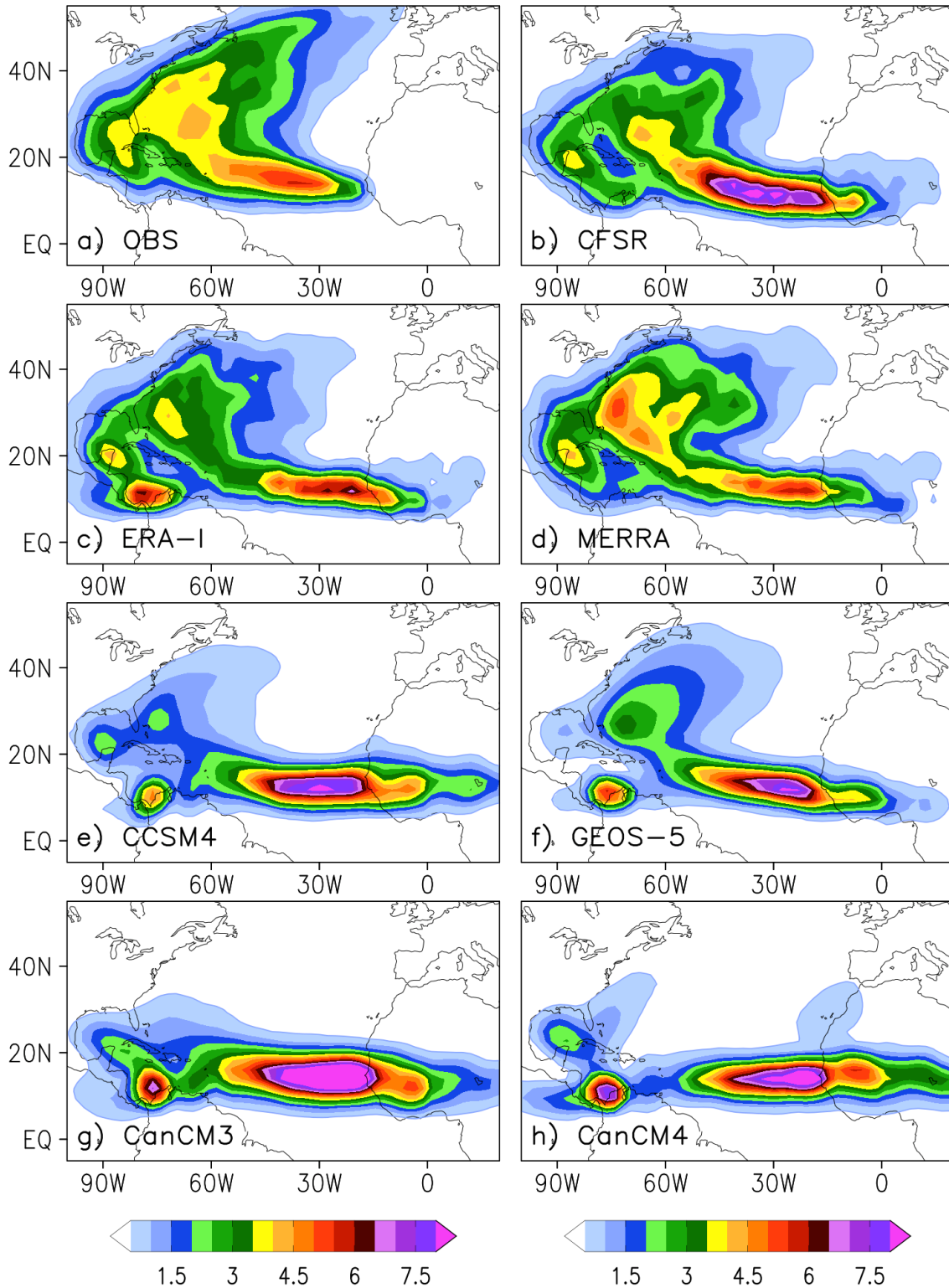


Figure 2: As in Fig. 1, but for the track density.

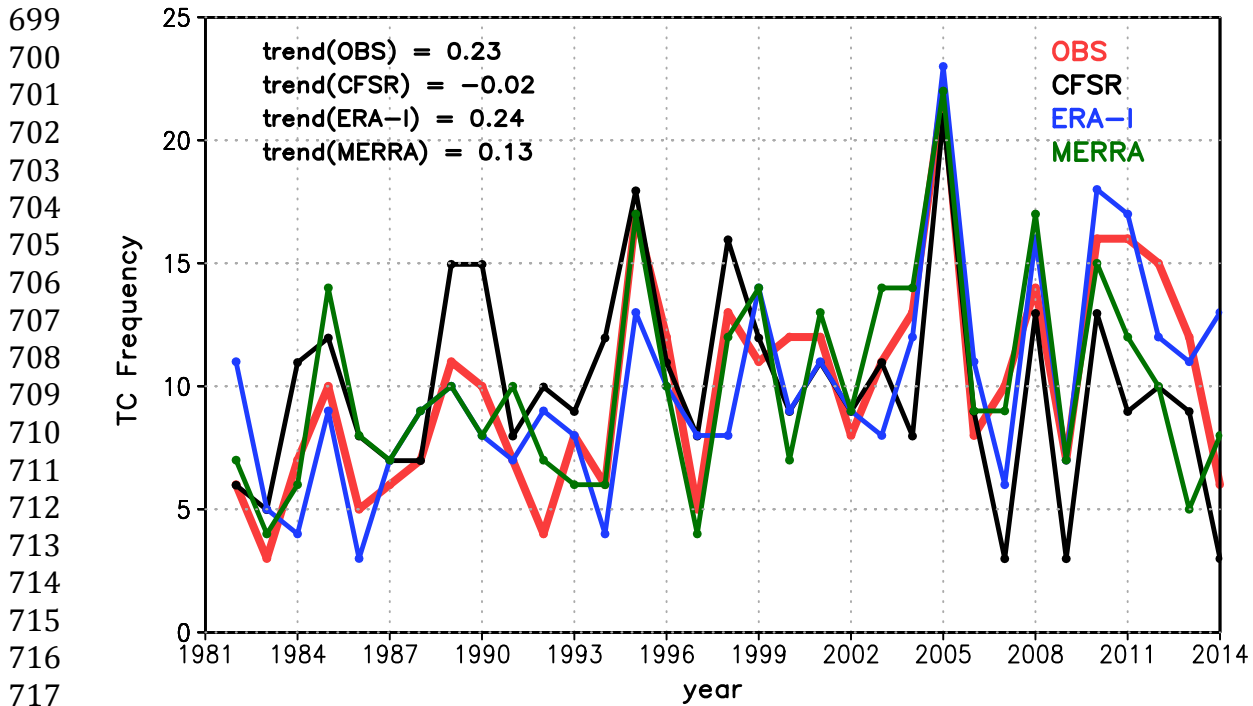


Figure 3: Time series of the NA MJJASON TC frequency based on the IBTrACS (OBS) data (red), and the CFSR (black), ERA-I (blue) and MERRA (green) reanalysis data sets. Linear trends for each time series are shown in the upper-left corner, units are counts per season per year.

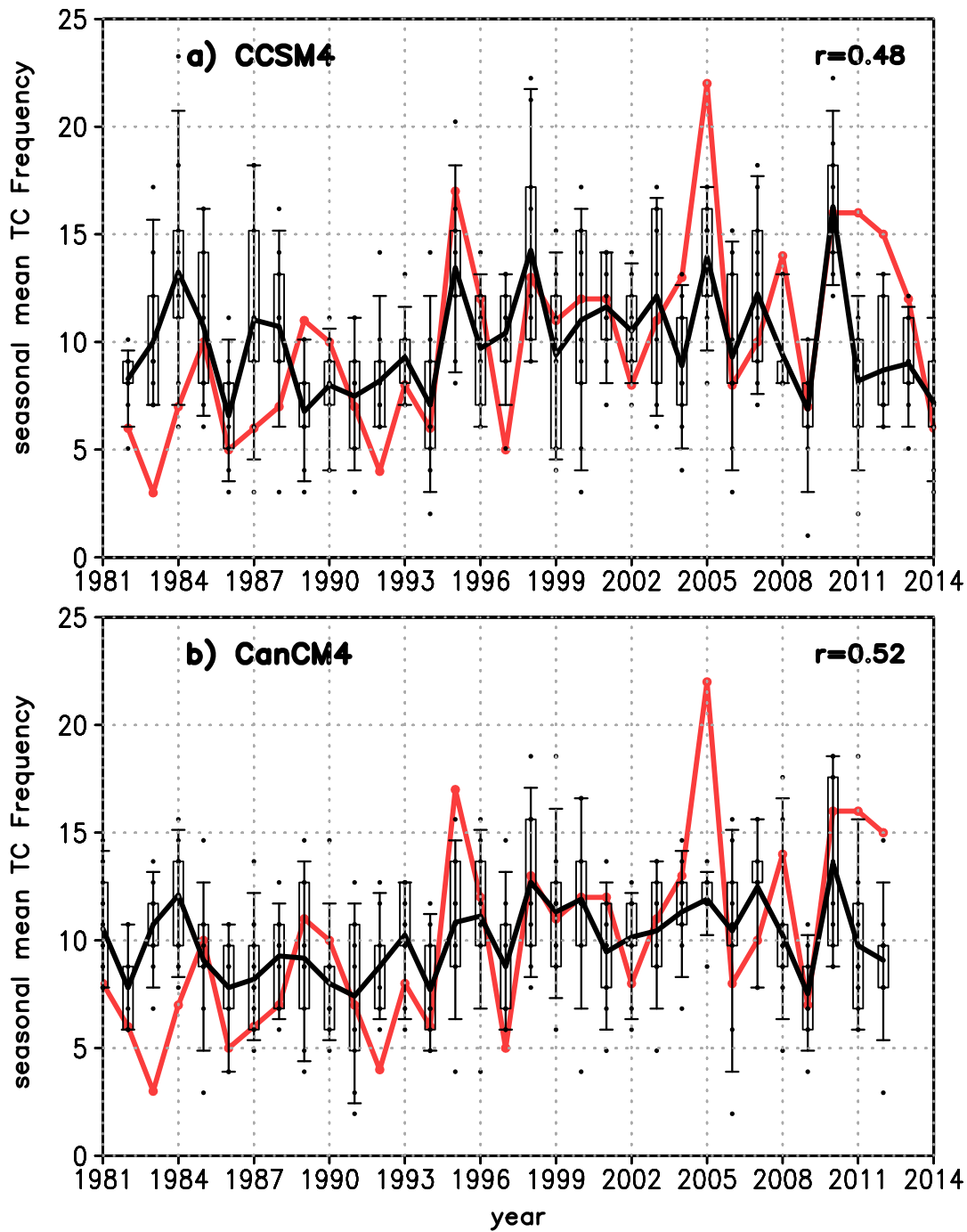
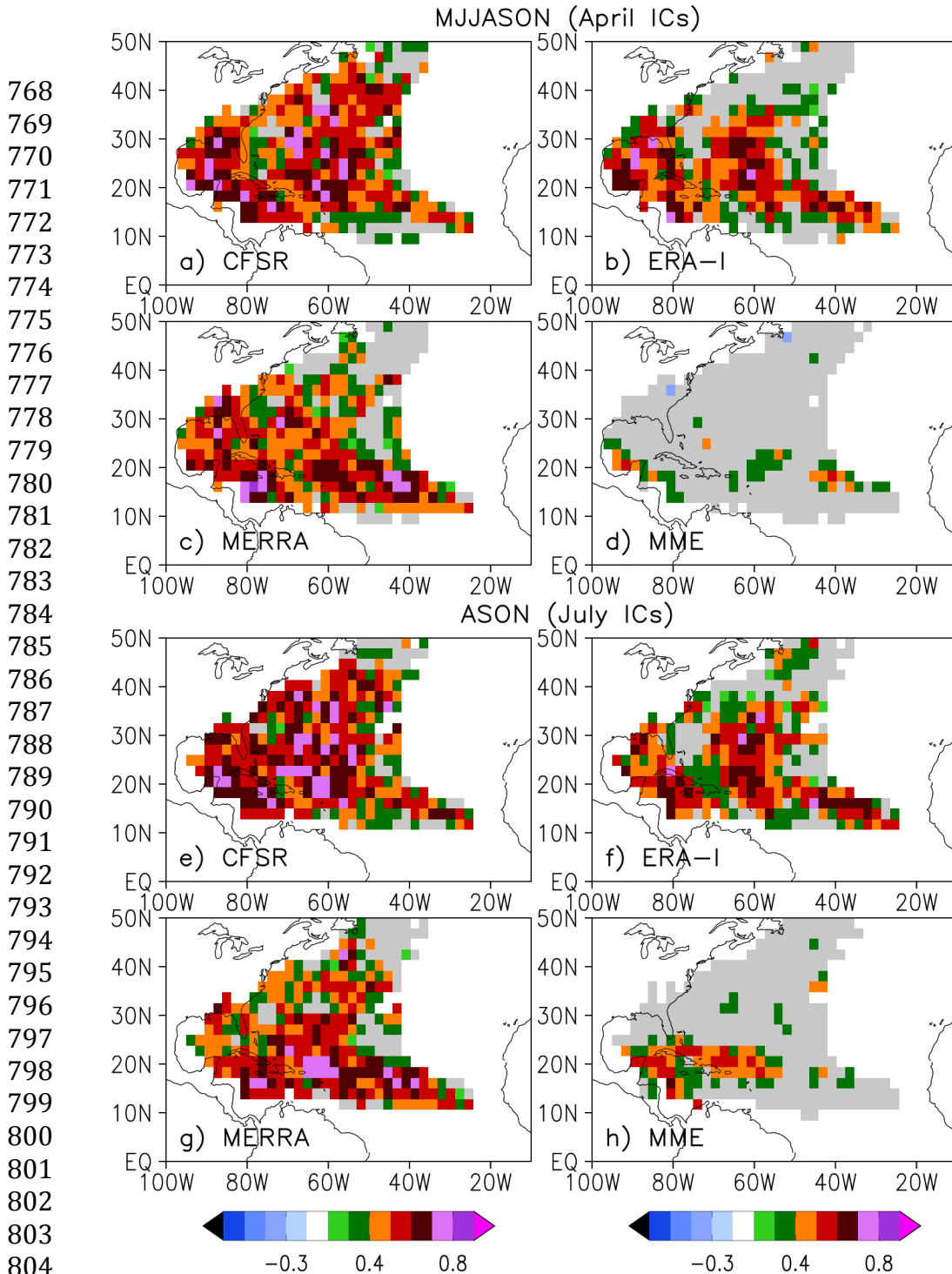


Figure 4: Retrospective forecasts (initialized in April) of the NA MJJASON TC frequency for the (a) CCSM4 and (b) CanCM4 NMME-Phase II models. Red and black lines show the observed time series and the ensemble-mean forecasts, respectively. Black dots mark predictions from the individual ensemble members. Box-and-whisker plots denote the 25th-75th and 10th-90th percentile ranges.



805 **Figure 5:** Rank correlation between the MJJASON observed (IBTrACS) and
806 reanalysis-derived TC track densities for 1982-2014 using (a) CFSR, (b) ERA-I, and
807 (c) MERRA. TC track density is defined as number density per season per unit area
808 equivalent to a 5° spherical cap. (E)-(g) are the same as (a)-(c) but for the ASON
809 season. (D) and (h) show retrospective rank correlation of the observed vs. MME
810 predicted TC track density for MJJASON (April ICs) and ASON (July ICs) of 1982-
811 2012, respectively. Values statistically significant at a two-sided $p=0.1$ level are
812 shown by color shading. Grey shading marks the regions where the observed track
813 density above zero for at least 25% of the years.

814
 815
 816
 817
 818
 819
 820
 821
 822
 823
 824
 825
 826
 827
 828
 829
 830
 831
 832
 833
 834
 835
 836
 837
 838
 839
 840
 841
 842
 843
 844
 845
 846
 847
 848
 849
 850
 851
 852
 853
 854

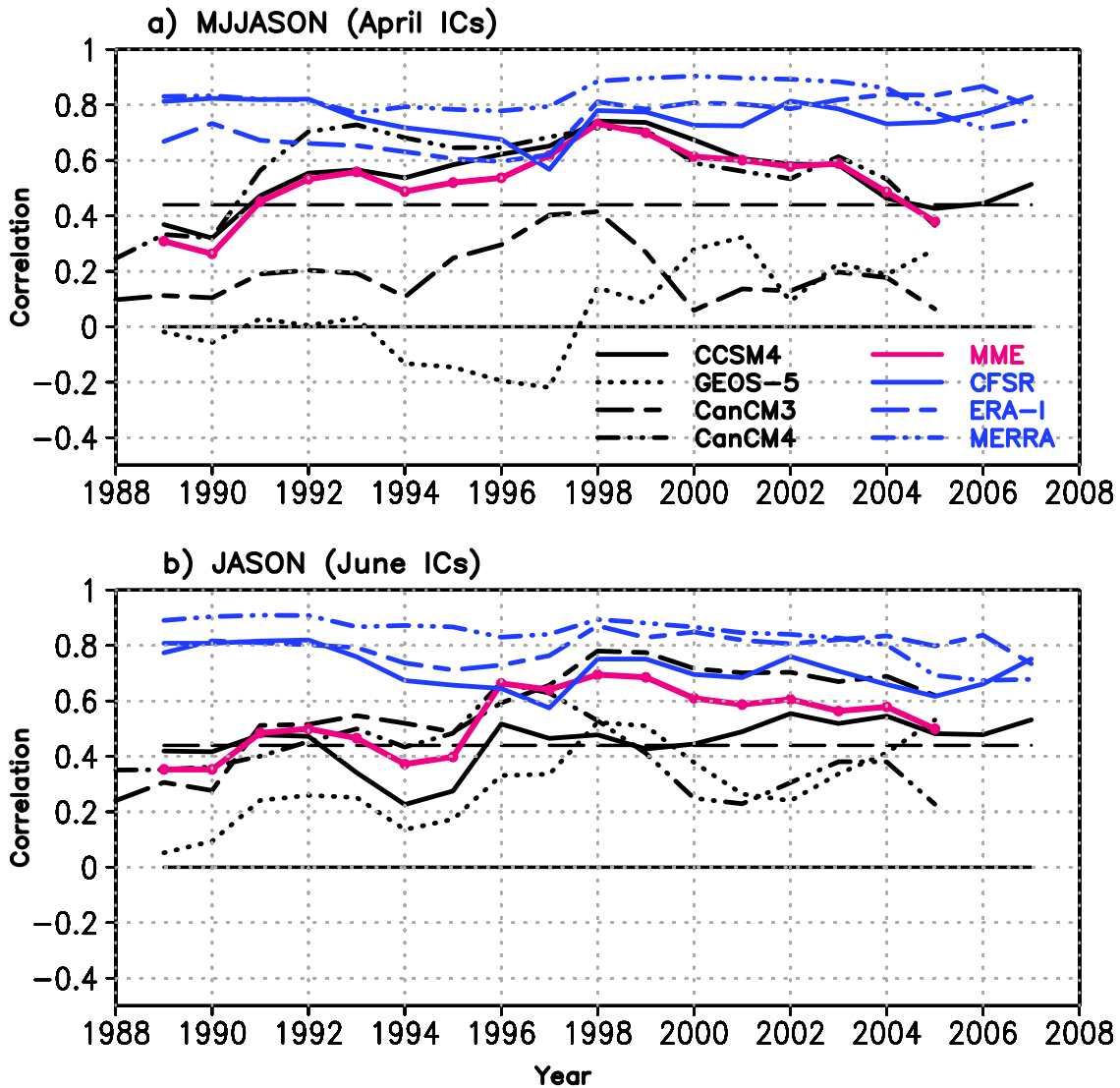
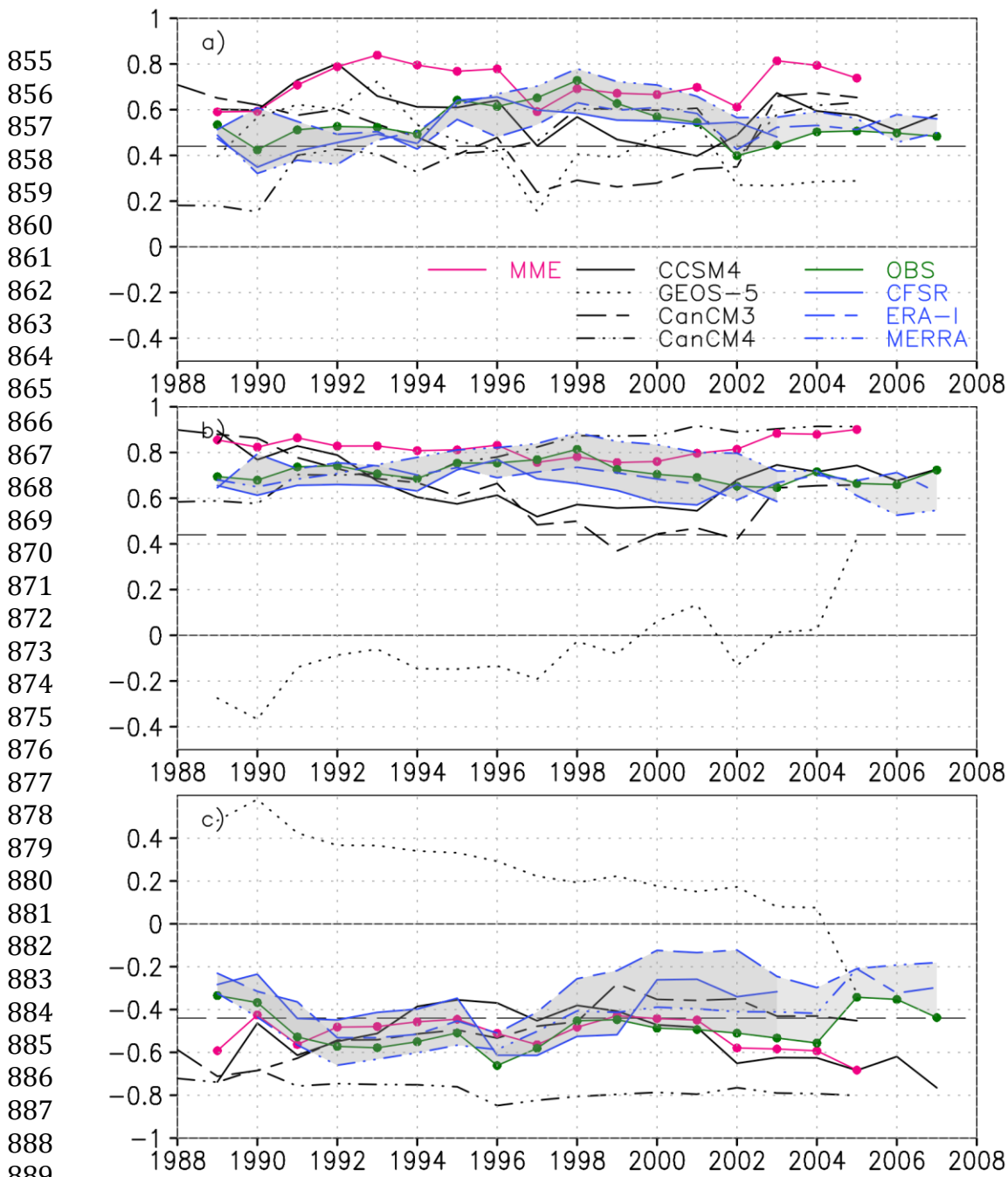


Figure 6: Sliding 15-year correlation of the predicted (ensemble mean) and reanalysis NA TC frequency with the observed (IBTrACS) for the (a) May-November season (forecasts initialized in April), and (b) July-November season (forecasts initialized in June). NMME-Phase II model results are shown in black and solid line for CCSM4, dotted for GEOS-5, long-dash-short-dash for CanCM3, and dot-dot-dash for CanCM4. Results for the MME mean are shown in magenta, and blue for the reanalyses (solid line for CFSR, long-dash-short-dash for ERA-I and dot-dot-dash for MERRA). Horizontal dashed line signifies statistically significant correlation. Horizontal axis marks the central year in the 15-year window.



890 **Figure 7:** Sliding 15-year correlation of the MJJASON NA TC frequency with the ASO
 891 mean (a) MDR SST index; (b) relative SST index; and (c) Niño-3.4 index (see
 892 definitions in the text) for observations (IBTrACS vs. OISSTv2), reanalysis and
 893 ensemble mean forecasts (initialized in April). NMME-Phase II model results are
 894 shown in black and solid line for CCSM4, dotted for GEOS-5, long-dash-short-dash
 895 for CanCM3, and dot-dot-dash for CanCM4. Results for the MME mean are shown in
 896 magenta, green for observations, and blue for the reanalyses (solid line for CFSR,
 897 long-dash-short-dash for ERA-I and dot-dot-dash for MERRA). Grey shading
 898 denotes the range of observed/reanalysis values. Horizontal dashed line signifies
 899 statistically significant correlation. Horizontal axis marks the central year in the 15-
 900 year window.

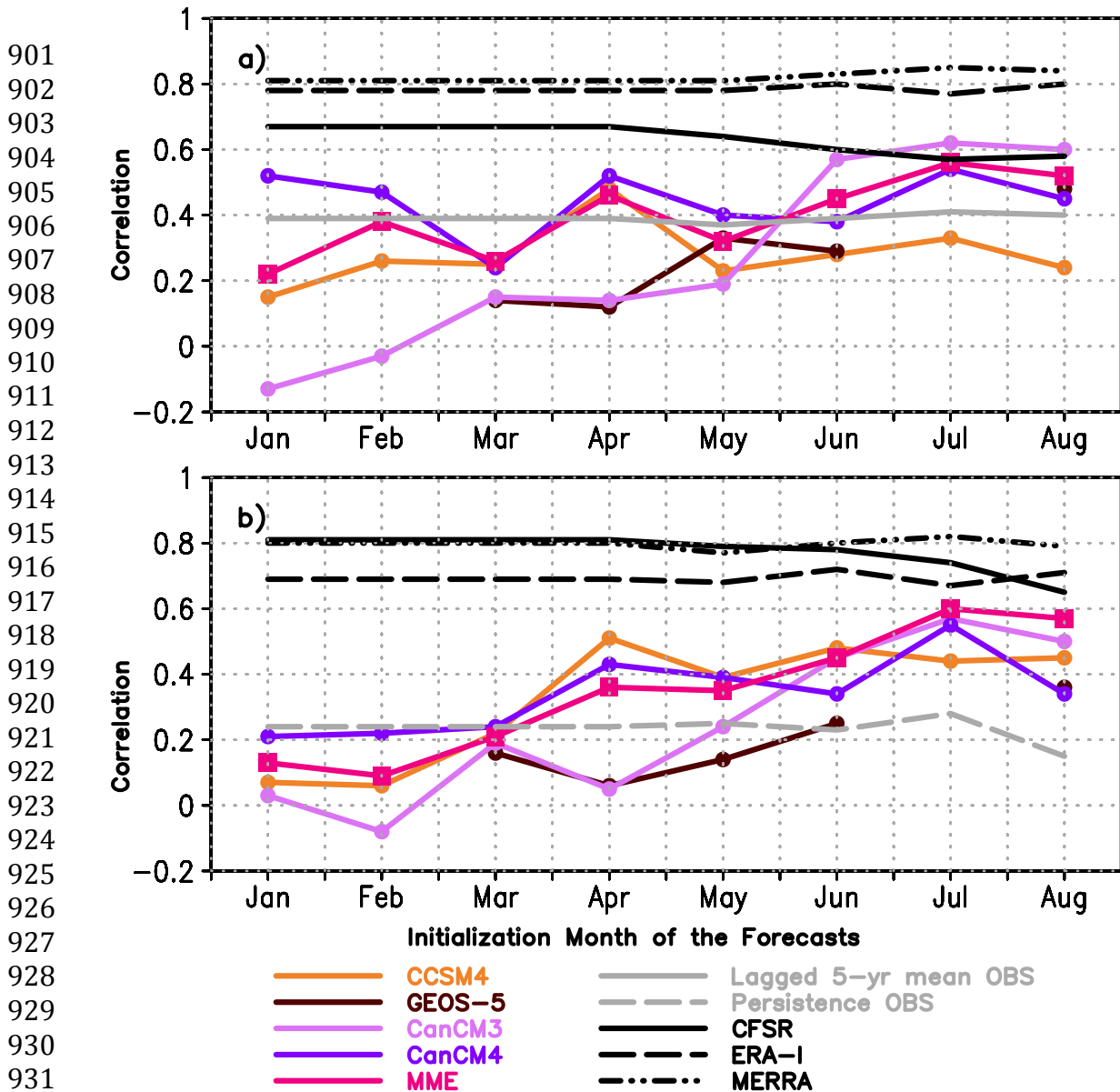


Figure 8: Correlation skill of the seasonal mean NA TC frequency for the NMME-Phase II models, the MME mean and the reanalyses as a function of forecast lead time, shown for the (a) full time series, and the (b) detrended time series. The solid colored lines display the skill of the CCSM4 (orange), GEOS-5 (brown), CanCM3 (lilac), CanCM4 (violet), and the MME mean (magenta). The black lines show the skill of CFSR (solid), ERA-I (long-dash), and MERRA (dot-dot-dash). Results shown are for the May-November average for forecasts initialized in January through April; June-November, July-November, August-November and September-November means when initialized in May, June, July and August, respectively. For the full time series, the skill is compared to a reference forecast comprising of the lagged 5-yr average of the observed TC frequency (solid gray; WMO 2008), and to persistence, or the previous season's observed TC frequency, (long-dash grey) for the detrended cases.

947
 948
 949
 950
 951
 952
 953
 954
 955
 956
 957
 958
 959
 960
 961
 962
 963
 964
 965
 966
 967
 968
 969
 970
 971
 972
 973
 974
 975

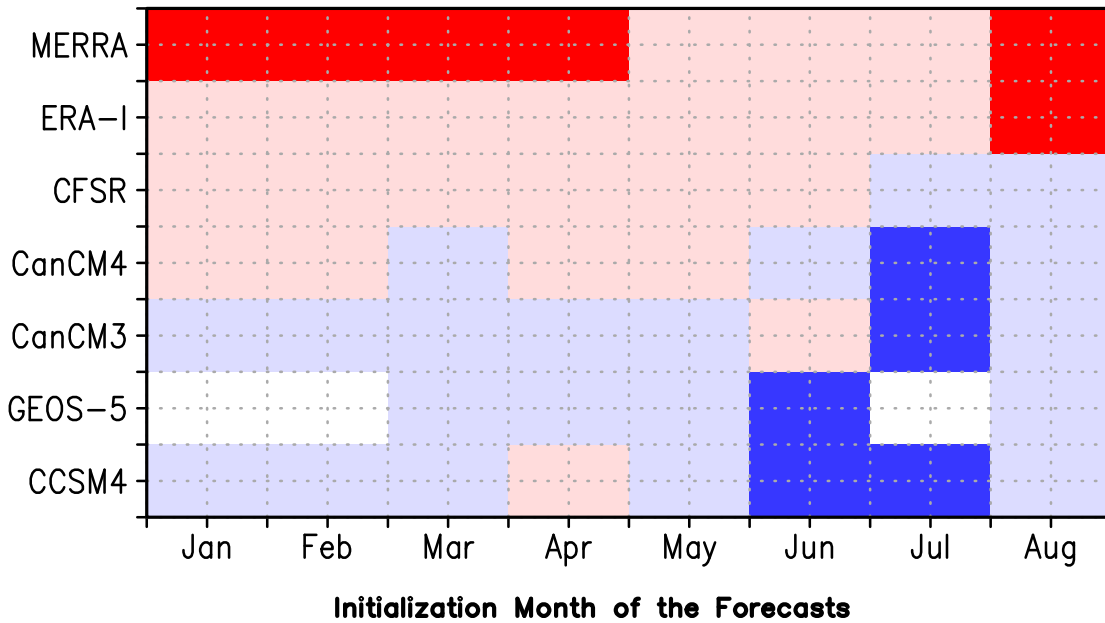


Figure 9: Difference between the squared error of the MME mean hindcasts and the squared error of the NMME-Phase II model or reanalysis indicated on the vertical axis, as a function of forecast lead time. Light blue (light red) color indicates that the MME mean squared error is smaller (larger) than the respective model/reanalysis. Dark blue (dark red) color indicates that the squared error of the MME mean is significantly smaller (larger) than the comparison model/reanalysis at the 95% confidence level using Wilcoxon signed-rank test. White blanks indicate that there are no results due to incompleteness/unavailability of the model data.

976

Table 1. NMME-Phase II models and forecasts.

Model Name	Modeling Center	Reference	Hindcast Period	Ensemble Size	Lead Times (months)	Atmospheric Model Resolution
CCSM4	University of Miami-Rosenstiel School for Marine and Atmospheric Science (UM-RSMAS)	Kirtman et al. (in prep.)	1982-2014	10	0-11	0.9x1.25 deg. L26
GEOS-5	National Aeronautics and Space Administration (NASA)	Verniers et al. (2012)	1982-2012	10	0-8	1x1.25 deg. L72
CanCM3	Canadian Centre for Climate Modeling and Analysis (CCCMA)	Merryfield et al. (2013)	1981-2012	10	0-11	T63L31
CanCM4	Canadian Centre for Climate Modeling and Analysis (CCCMA)	Merryfield et al. (2013)	1981-2012	10	0-11	T63L35

977

978

979 **Table 2.** Linear correlation of the predicted (ensemble mean) and reanalysis NA TC frequency with the observed
 980 (IBTrACS) for 1982-2014 for the reanalyses data sets, and the time periods listed in Table 1 for the forecasts. Results
 981 are shown for May-November (MJJASON), August-November (ASON) and September-November (SON) seasons with
 982 forecasts initialized in April, July and August, respectively. Multi-model ensemble mean (MME) is based on four or
 983 three models listed depending on data availability, as indicated. Values in parentheses show correlation coefficients
 984 computed for the detrended time series. Boldface marks values that are statistically significant at the 95% confidence
 985 level.

Season (ICs)	CCSM4	GEOS-5	CanCM3	CanCM4	MME	CFSR	ERA-I	MERRA
MJJASON (April ICs)	0.48 (0.51)	0.12 (0.06)	0.14 (0.05)	0.52 (0.43)	0.46 (0.36)	0.67 (0.81)	0.78 (0.69)	0.81 (0.80)
ASON (July ICs)	0.33 (0.44)	-*	0.62 (0.57)	0.54 (0.55)	0.56 (0.60)	0.57 (0.74)	0.77 (0.67)	0.85 (0.82)
SON (August ICs)	0.24 (0.45)	0.48 (0.36)	0.60 (0.50)	0.45 (0.34)	0.52 (0.57)	0.58 (0.65)	0.80 (0.71)	0.84 (0.79)

986

987 -* incomplete data

988

989 **Table 3.** RMSE between the calibrated ensemble-mean forecasts and the observations (IBTrACS) of the NA TC
 990 frequency based on the time periods listed in Table 1, and between the reanalyses and observed NA TC frequency for
 991 1982-2014. Results are shown for May-November (MJJASON), August-November (ASON) and September-November
 992 (SON) seasons with forecasts initialized in April, July and August, respectively. Multi-model ensemble mean (MME) is
 993 based on four or three models listed depending on data availability, as indicated. Values in parentheses show RMSE for
 994 the detrended time series.

Season (ICs)	CCSM4	GEOS-5	CanCM3	CanCM4	MME	CFSR	ERA-I	MERRA
MJJASON (April ICs)	3.73 (3.15)	4.32 (3.54)	4.27 (3.58)	3.66 (3.06)	3.87 (3.18)	3.37 (2.37)	2.81 (2.80)	2.57 (2.40)
ASON (July ICs)	3.73 (3.05)	-*	2.89 (2.39)	3.09 (2.44)	3.09 (2.28)	3.34 (2.46)	2.44 (2.43)	1.95 (1.84)
SON (August ICs)	2.93 (2.25)	2.61 (2.23)	2.32 (2.09)	2.59 (2.30)	2.56 (2.02)	2.42 (2.01)	1.79 (1.78)	1.57 (1.54)

995

996 -* incomplete data

997

998 **Table 4.** The SPRvERR for the calibrated predicted NA TC frequency based on the time periods listed in Table 1.
 999 Results are shown for May-November (MJJASON), August-November (ASON) and September-November (SON) seasons
 1000 with forecasts initialized in April, July and August, respectively. Multi-model ensemble mean (MME) is based on four or
 1001 three models listed depending on data availability, as indicated. Values in parentheses show SPRvERR for the
 1002 detrended time series.

Season (ICs)	CCSM4	GEOS-5	CanCM3	CanCM4	MME
MJJASON (April ICs)	0.79 (0.91)	0.59 (0.70)	0.60 (0.69)	0.74 (0.86)	0.74 (0.88)
ASON (July ICs)	0.74 (0.88)	-*	0.93 (1.07)	0.93 (1.11)	1.00 (1.31)
SON (August ICs)	0.75 (0.93)	0.77 (0.88)	0.96 (1.04)	0.90 (0.99)	0.97 (1.20)

1009

1010 -* incomplete data

1011

1012 **Table 5.** As in Table 2 but for TC days. Only values for the full time series are shown.

1013

Season (ICs)	CCSM4	GEOS-5	CanCM3	CanCM4	MME	CFSR	ERA-I	MERRA
MJJASON (April ICs)	0.39	0.21	0.29	0.57	0.46	0.85	0.82	0.82
ASON (July ICs)	0.37	-*	0.67	0.55	0.59	0.80	0.82	0.83
SON (August ICs)	0.37	0.54	0.66	0.38	0.59	0.76	0.80	0.79

1014

1015 -* incomplete data

1016