Climate Dynamics (2018) 51:2725–2742
https://doi.org/10.1007/s00382-017-4040-z

Potential predictability and forecast skill in ensemble climate forecast: a skill-persistence rule

Yishuai Jin[1] · **Xinyao Rong**[2] · **Zhengyu Liu**[3]

Received: 19 March 2017 / Accepted: 6 December 2017 / Published online: 14 December 2017

Abstract
This study investigates the factors relationship between the forecast skills for the real world (actual skill) and perfect model (perfect skill) in ensemble climate model forecast with a series of fully coupled general circulation model forecast experiments. It is found that the actual skill for sea surface temperature (SST) in seasonal forecast is substantially higher than the perfect skill on a large part of the tropical oceans, especially the tropical Indian Ocean and the central-eastern Pacific Ocean. The higher actual skill is found to be related to the higher observational SST persistence, suggesting a skill-persistence rule: *a higher SST persistence in the real world than in the model could overwhelm the model bias to produce a higher forecast skill for the real world than for the perfect model*. The relation between forecast skill and persistence is further proved using a first-order autoregressive model (AR1) analytically for theoretical solutions and numerically for analogue experiments. The AR1 model study shows that the skill-persistence rule is strictly valid in the case of infinite ensemble size, but could be distorted by sampling errors and non-AR1 processes. This study suggests that the so called "perfect skill" is model dependent and cannot serve as an accurate estimate of the true upper limit of real world prediction skill, unless the model can capture at least the persistence property of the observation.

Keywords Predictability · Seasonal forecast · Perfect model · CGCM · AR1 model

1 Introduction

For climate predictions of seasonal to decadal time scales, one approach to assess the potential predictability is to perform ensemble climate forecasts in dynamic models in the "perfect model" framework (Griffies and Bryan 1997; Chen et al. 2010; Dunstone and Smith 2010; Teng et al. 2011; Sévellec and Fedorov 2013). The perfect model forecast skill (perfect skill hereafter) has often been treated as the benchmark for the forecast skill for the real world climate prediction (actual skill hereafter). Perfect skill is often thought to be the upper limit of the actual skill such that the gap between the perfect skill and actual skill is often considered as the "room for improvement", i.e., the gap by which the actual skill can be enhanced (Boer et al. 2013; Younas and Tang 2013; Holland et al. 2013; Becker et al. 2013, 2014). For example, Becker et al. (2014) define the potential predictability by the perfect skill, namely, the skill by verifying model's ensemble mean, based on N-1 members, against the one member that is left out. They found that in most models the potential predictability is substantially higher than the forecast skill, thus "providing some hope for improved SST prediction". However, a few studies have recently noted examples of lower perfect skill than actual skill (Mehta et al. 2000; Kumar et al. 2014). Kumar et al. (2014) studied the relationship between the perfect and actual skills in two operational seasonal forecast models from a signal-to-noise ratio perspective with the focus on the anomaly correlation coefficient (ACC) skill. They suggested that, a priori, there is not necessarily a relation between the perfect skill and the actual skill. The actual skill can be higher than the perfect skill in some region because of the different statistical properties

✉ Xinyao Rong
rongxy@cma.gov.cn

✉ Zhengyu Liu
liu.7022@osu.edu

[1] Department of Atmospheric and Oceanic Sciences, School of Physics, Peking University, Beijing, China

[2] State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China

[3] Atmospheric Science Program, Department of Geography, Ohio State University, Columbus, OH 43210, USA

of the forecast variability from the observation. In particular, the anomalous temporal correlation of the forecast seems to serve as a useful indicator of the discrepancy between actual and perfect skills, with a higher correlation in forecasts favoring a higher actual skill. In fact, using the difference of persistence to indicate difference of skill has previously been applied for subseasonal prediction (Pegion and Sardeshmukh 2011). However, it is unclear if the relation between the persistence and forecast skill is necessary, and the relation between persistence and forecast skill changes for different forecast and different models.

Here, we extend the work of Kumar et al. (2014) by studying the factors contributing to the difference between the actual and perfect forecast skills in terms of both ACC and root-mean-square error (RMSE), systematically. We will study the forecast skill for seasonal climate forecasts in a fully coupled general circulation model (CGCM). The difference between actual skill and perfect skill is further studied using simple statistical models. In particular, our study in the first order regressive process (AR1) model suggests a rule for predicting the difference between the actual and perfect skills using the difference of the persistence between the observation and the forecast, or the skill-persistence rule: a higher persistence in the real world will lead to a higher actual skill, and vice versa. This rule is found to be able to explain the CGCM results with reasonable success, especially for the case of higher model persistence. Different from the traditional approach by Kumar et al. (2014) and Becker et al. (2014), where the perfect skill is evaluated by taking one member of the ensemble forecast as the truth, in the present study the perfect skill is evaluated by a continuous simulation of CGCM. Namely, the CGCM is first integrated to generate a continues time series (i.e., the "truth" in perfect model framework) and the same model is used to predict this time series. The reason is that the traditional strategy using prediction to calculate the persistence may degrade the persistence due to "initial shock" (e.g., Zhang et al. 2007; Pohlmann et al. 2016), which will be avoided in our continuous simulation such that the perfect skill of the model can be compared more directly against the actual skill for the observation. The difference between our strategy and the traditional approach, nevertheless, is minor as will also be discussed.

The rest of this paper is arranged as follows. Section 2 describes the model, experiments and measures of forecast skill. The CGCM seasonal forecast skill and its connection with persistence are presented in Sect. 3. In Sect. 4, we discuss the result of statistical model forecasts. A summary and discussion are given in Sect. 5.

2 Model, experiment and analysis methods

2.1 Model and experiment

The CGCM used in this study is the Fast Ocean Atmosphere Model (FOAM, Jacob 1997; Tobis et al. 1997). The atmospheric component consists of a spectral dynamical core of an R15 resolution (approximate to 7.5° longitude, 4° latitude, 18 vertical levels), and physics from the CCM3 model. The ocean component is a z-coordinate model stemmed from the GFDL MOM1.0 with a resolution of 2.8° longitude, 1.4° latitude and 24 vertical layers. A thermodynamic sea ice model with the same grid set as the ocean is adopted for the sea ice component. FOAM has shown comparable performance with current CGCMs on ENSO (Liu et al. 2000) and Pacific Decadal Oscillation (PDO) (Wu et al. 2003; Liu et al. 2007), and has been used for ensemble coupled data assimilation for the study of parameter estimation (Liu et al. 2014) and climate dynamics (Lu et al. 2016).

We first perform a control run by integrating FOAM from a rest ocean of observed climatological temperature and salinity for 500 years with the constant pre-industrial atmospheric CO_2 level. From the end of the control run, a historical run (HIST) similar to the ones in CMIP5 (Taylor et al. 2012) is conducted from year 1871 to 2005. The greenhouse gases used in HIST are the CO_2 equivalence concentration (CO_2EQ) which aggregates all anthropogenic forcing including CO_2, CH_4, N_2O, aerosol, ozone etc. (Meinshausen and Smith 2011). To compare the actual skill and perfect skill, two parallel forecast experiments with identical assimilation schemes are performed: the real world forecast (RW) and the perfect model forecast (PM). The RW resembles the operational forecast, with the monthly SST anomalies from the HadISST (Rayner et al. 2003) assimilated into FOAM for the initialization of the forecasts, and the forecast skill is evaluated against the observational data (HadISST). The PM takes the HIST run as the "truth", and an artificial "observation" is constructed by adding small perturbations (random Gaussian white noise) onto the "truth" SST time series. The constructed "observation" (i.e., the perturbed SSTs) is subsequently assimilated into an independent FOAM simulation to initialize the PM forecasts, and correspondingly the forecast skill is evaluated against the known "truth". Hereafter, the forecast skill of RW is referred as the actual skill since in RW we use the model to predict the real world, and the skill of PM is called the perfect skill as in PM the model is used to predict the state generated by the identical model.

The data assimilation scheme for the model initialization in RW and PM forecasts is the ensemble adjustment Kalman filter (EAKF), which is a specific type of square root filter algorithm of ensemble Kalman Filter (Anderson 2001, 2003). We use an ensemble size of 20 for our experiments.

The assimilation starts from the 1st of January of 1950, with identical initial conditions for ocean, land and sea ice but perturbed atmosphere. The observations are monthly SST covering all grid points of the global ocean. The HadISST dataset is remapped onto FOAM's grid by a bilinear interpolation. The observational error of the SST in RW (i.e., HadISST) and PM (i.e., the random noise added onto the "truth") is assumed to be a Gaussian white noise with a standard deviation of 0.2 °C, about half of global mean SST trend from 1960 to 2005. Using the cross-covariances between temperatures and salinities, the SST observations are utilized to update the ocean temperature and salinity above 400 m. A horizontal localization scheme of Gaspari and Cohn (1999) is adopted with the influence radius of three zonal grid intervals. The assimilation takes place at the end of each month. The monthly mean values are used to calculate the increments of state variables, which are subsequently added to the states at the first step of next month. The assimilation result shows that the assimilated SST remains close to the observations. The analysis error of monthly SST is less than 0.2 °C in most regions, with similar patterns and magnitudes between PM and RW experiments.

The forecasts in RW and PM start from the 1st day of each month from 1960 to 2005, containing 552 forecasts with an ensemble size of 20 and forecast length of 12 months. The initial conditions for the atmosphere, land, ocean and sea ice in RW and PM forecasts are taken from the assimilated simulation of corresponding starting time. While the atmosphere, land, sea ice states are not directly assimilated, the observational information may be transferred from the ocean into other components as the SST is assimilated in a coupled system. Analogous to the decadal prediction scheme of CMIP5, the CO2EQ in the above forecasts is prescribed as in the HIST run.

2.2 Verification measures

The forecast skill will be verified in two metrics:the anomaly correlation coefficient (ACC) and the root-mean-square error (RMSE). ACC is defined as the temporal correlation coefficient between anomalies of the ensemble mean forecast and the corresponding "truth":

$$\mathrm{ACC} = \frac{\langle F_i, O_i \rangle}{\sqrt{\langle F_i, F_i \rangle \cdot \langle O_i, O_i \rangle}}, \tag{1}$$

where F_i is the ensemble mean forecast anomaly for forecast month i (i.e., 552 forecasts for each month), O_i is the verifying observed anomaly, and "$< >$" denotes the variance over all the months in verifying time series. For PM forecast O_i is the HIST simulation (i.e., the "truth" in PM), while O_i for RW forecast is the HadISST anomalies.

The normalized RMSE is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{\langle F_i - O_i, F_i - O_i \rangle}{\langle O_i, O_i \rangle}}, \tag{2}$$

where F_i and O_i are of the same meaning as in ACC.

3 Forecast skill in CGCM

We first compare the forecast skill between the RW and PM forecasts in FOAM, and investigate the cause for the difference of the forecast skill between RW and PM. For convenience, we will refer the RW skill and PM skill hereafter as the actual skill and perfect skill, respectively.

3.1 Forecast skill in RW and PM forecasts

Figure 1 shows the ACC of SST for different forecast lead times in PM and RW forecasts respectively. Similar to previous studies (Kumar et al. 2014), high ACC are primarily located in the tropical oceans, especially the central-eastern equatorial Pacific, indicating the dominant role of ENSO in the predictability of seasonal to interannual forecasts. Several differences can be found between PM and RW forecast skills. First, the maximum ACC of PM in the tropical Pacific is shifted to the west in comparison with RW, as a result of the excessive westward penetration of the cold tongue and associated westward displacement of SST variability in FOAM (Jacob 1997; Liu et al. 2003). Second, over the tropical oceans, the ACC in RW is overall higher than PM, reminiscent of the result of Kumar et al. (2014) where the actual skill on a portion of grid points is higher than the perfect skill. Third, over most areas of the middle and high latitudes the ACC in RW is lower than PM, particularly on the North Atlantic and the Antarctic Circumpolar Current area, which seems to be consistent with the view that the perfect skill may serve as the upper bound of the actual skill in the presence of the model bias. Similar features can be also found in RMSE in Fig. 2, indicating that the skill measured by ACC and RMSE is consistent.

The differences of the SST ACC and RMSE between PM and RW forecasts are further shown in the global maps of their differences (Fig. 3). Consistent with the discussions on Figs. 1 and 2, the actual skill can exceed the perfect skill over large areas, notably in the tropical central to eastern Pacific and Indian Ocean (blue shaded area). In addition, the difference between actual and perfect skill becomes larger as forecast time increases. One noticeable asymmetry feature between the positive and negative ACC differences is that on average (Fig. 3a–c) the positive ACC difference over the southern high latitude ocean and the North Pacific can reach + 0.4 to + 0.6, considerably larger than the magnitude of the negative ACC difference in the tropical Pacific and

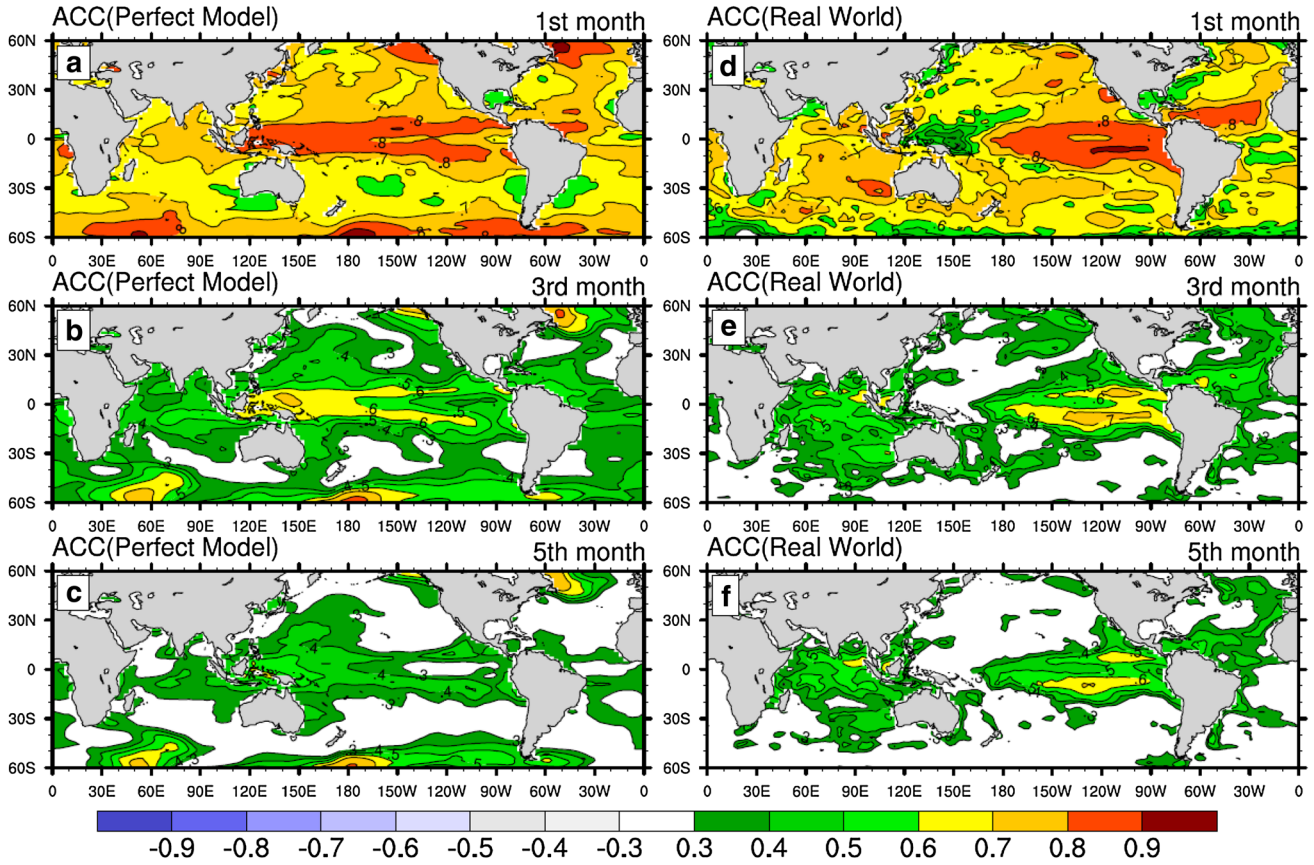

Fig. 1 ACC of real world (right) and perfect model (left) seasonal forecasts for different lead times. From top to bottom: ACC for leads of 1, 3 and 5 months

Indian Ocean of −0.2 to −0.3. As will be discussed later, this asymmetry is attributed to the model bias as well as the different predictability between the real world and FOAM model. Another reason for the asymmetry is that, over the tropical ocean the ACC is already high closer to the generic upper bound 1, leaving less room for the ACC difference.

We also calculate the differences of the SST ACC and RMSE between PM and RW forecasts using the traditional perfect model approach, in which one member of the ensemble forecast member is used as the truth (e.g., Kumar et al. 2014; Becker et al. 2014) (figure not shown). The results show similar features to those of our approach. Lower perfect skill is found over the tropical oceans, including the tropical Indian Ocean and the eastern equatorial Pacific Ocean. The spatial correlation coefficients of the ACC (RMSE) differences global maps between two methods for forecast months 1, 3 and 5 are 0.40(0.60), 0.64(0.89) and 0.67(0.91), respectively, indicating that the perfect skill calculated by two methods are essentially same. Note that the correlation coefficients at the first forecast month are substantially lower than those of longer forecast times. This is caused by the degraded persistence due to "initial shock" in the traditional method.

To examine the time evolution of the actual skill and perfect skill in different regions, we plot the time series of regional averaged ACC and RMSE over three representative regions (Fig. 4). The first two regions (tropical Indian Ocean and the Nino3.4 region) have the actual skill higher than the perfect skill, while the third region (North Pacific) has a lower actual skill. While both forecast skill decreases with forecast time, the forecast skill first separate from each other with the maximum occurring at about lead months 5–6, and eventually converge, after about lead month 10–12. It is speculated that, with the further increase of the forecast lead time, the model bias will gradually dominate the forecast error, leading to a lower actual skill than perfect skill, then eventually both skill converges toward zero and the differences will be undistinguished. For the North Pacific region, the opposite happens (Fig. 4c, f), where the perfect skill is substantially higher than the actual skill for all forecast months and remains high at long lead times.

The evolution of the forecast skill can be seen more clearly in the evolution of the difference of forecast skill (black lines). Here, the significance of the differences in forecast skill is tested with a Montel Carlo approach by randomly scrambling the forecast anomalies of the total 552

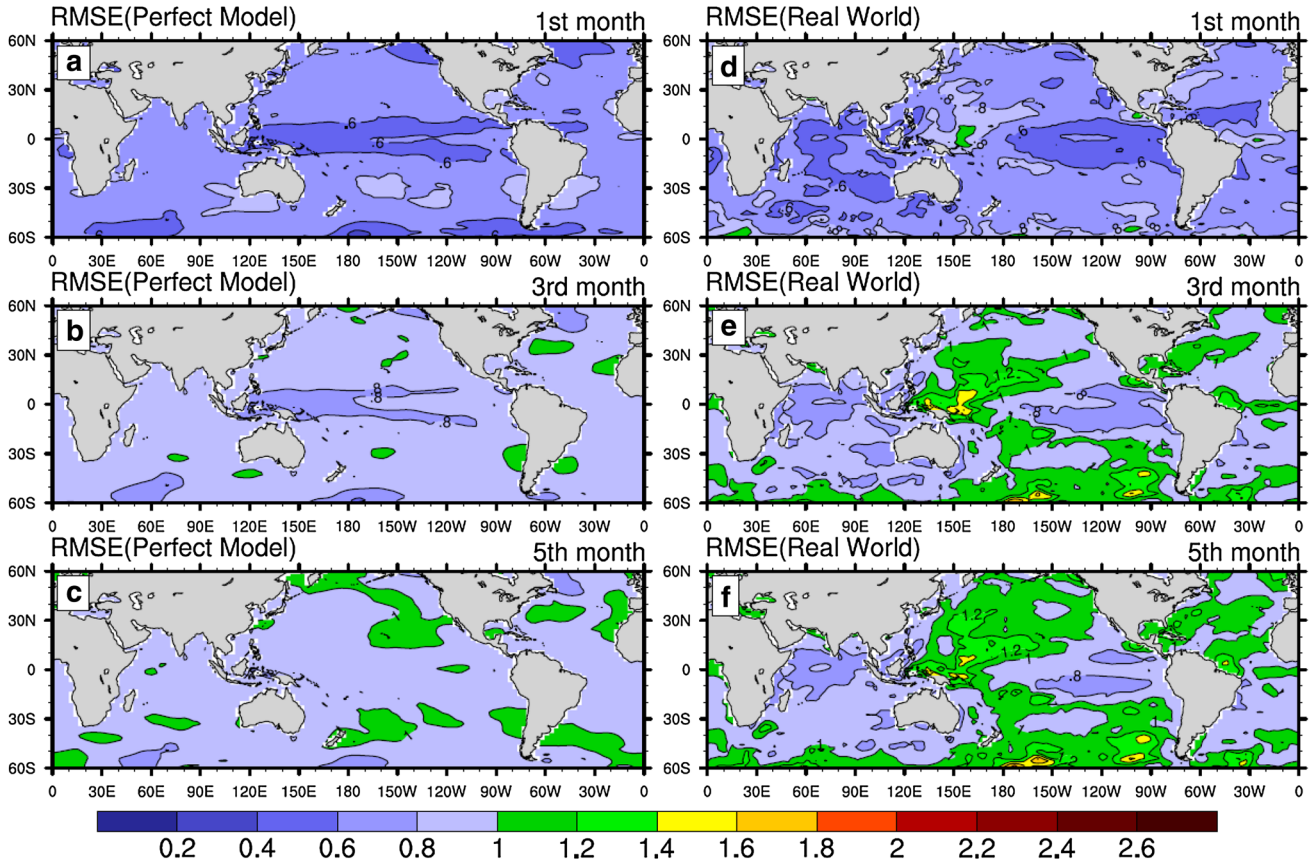

Fig. 2 Same as Fig. 1 but for RMSE. The RMSE of real world and perfect model forecasts have been normalized by the SST standard deviation of HadISST and HIST run, respectively

forecasts at each lead month and then calculating the differences of ACC and RMSE repeatedly for 1000 times. Indeed, the actual skill is generally higher than the perfect skill in the tropical Indian Ocean and Nino3.4 areas and the differences exceed the 5% significance level (grey shade bar) for most forecast months (month 2–9). The differences of skill reach the maximum at about 5–6 months and vanishing eventually after 10–12 months. In the North Pacific the perfect skill is significantly higher than the actual skill. We also calculate the forecast skill initialized from different calendar months (not shown). It is found that the higher actual skill in the tropical Indian Ocean and Nino3.4 areas, as well as the lower actual skill in the North Pacific, is not very sensitive to the initial months of forecast.

3.2 Cause of the difference in actual skill and potential skill

Why the actual skill can exceed the perfect skill, given the inevitable bias in the model? Kumar et al. (2014) argue that this is related to the difference of persistence (autocorrelation) between forecast model and real world, and a higher anomaly temporal autocorrelation in real world than model tends to favor a higher actual skill. Moreover, a higher (lower) persistence also indicates lower (higher) forecast dispersion. However, their persistence is calculated using the forecast anomalies among different forecast lead times and they only compared the forecast persistence characteristics between two seasonal prediction systems, but not directly against the observations, which is calculated on a single time series. As such, they did not compare the autocorrelations between the real world variability and model variability directly. Here, to examine the role of the difference of statistical properties between the real world and model on the different forecast skill more clearly, we will compute the persistence in the real world and the model consistently on the time series of the observational data for RW (HadISST) and the "truth" run (HIST) for PM. Figure 5 shows the difference of autocorrelation between real world (HadISST) and HIST for different lagged months, which is meaningful because both autocorrelations are calculated on the respective time series the same way. Comparing Fig. 5 with Fig. 3, the relation between forecast skill and persistence emerges. In general, regions with higher RW persistence (blue shaded

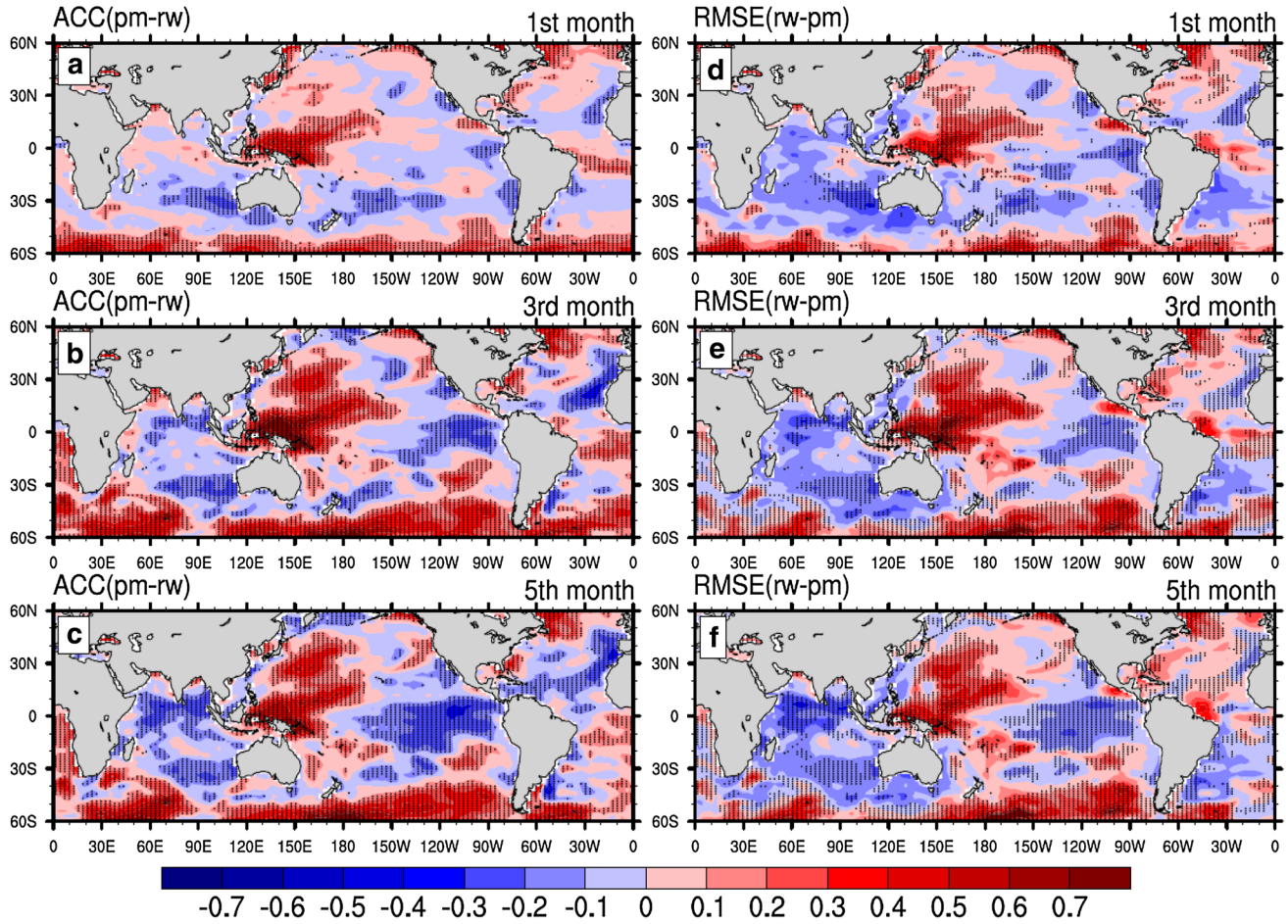

Fig. 3 Differences of ACC (left, PM minus RW) and RMSE (right, RW minus PM) between the real world and perfect model seasonal forecasts. From top to bottom: leads of 1, 3 and 5 months. The points above 5% significance levels are dotted

areas in Fig. 5), for example, the tropical Indian Ocean and the equatorial Pacific, tend to overlap with regions of higher actual skill than the perfect skill (blue shaded areas in Fig. 3). The difference of forecast skill becomes larger as the difference of persistence increases with lead times. Over the middle and high latitude oceans, the persistence of the model is close to (the North Pacific) or even exceeds (red shaded areas in the Southern Ocean) the real world, therefore a lower actual skill appears over there. The reason is that, even real world and FOAM are of same persistence, model bias will also lead to a lower actual skill. Since the persistence of FOAM is higher than real world in the middle and high latitudes, the model bias in RW and the higher persistence in PM will be combined to cause an even higher perfect skill there.

The analysis of the FOAM forecast seems to suggest a skill-persistence rule, which uses the difference of the persistence to predict the difference of the forecast skill, as follows: a larger persistence in the real world than in the model infers a higher actual skill than perfect skill, and vice versa. This rule, as will be seen in the next section, can be shown strictly valid for a theoretical first order regressive (AR1) process. This rule can explain to certain extent the results of our complex CGCM forecasts, but does not hold completely, as will be discussed below.

The relationship between the forecast skill and persistence can be studied more quantitatively in a scatter diagram (Fig. 6), which shows the relation between the difference of forecast skill (y-axis, ACC in a–c and RMSE in d–f) and the difference of persistence (x-axis), for the forecast months of 1, 3 and 5. Here the persistence corresponding to forecast months 1, 3 and 5 is calculated as the lagged 1, 3 and 5 months' autocorrelation, respectively. As the y-axis for ACC is (PM-RW) and for RMSE is (RW-PM), a more positive (negative) value in x coordinates indicates a higher (lower) persistence in model relative to the real world, and a more positive (negative) value on y coordinates corresponding to a higher (lower) perfect skill than the actual skill. The percentage of grid points in each quadrant relative to the total grid points is also labeled in each panel. First of all, it is

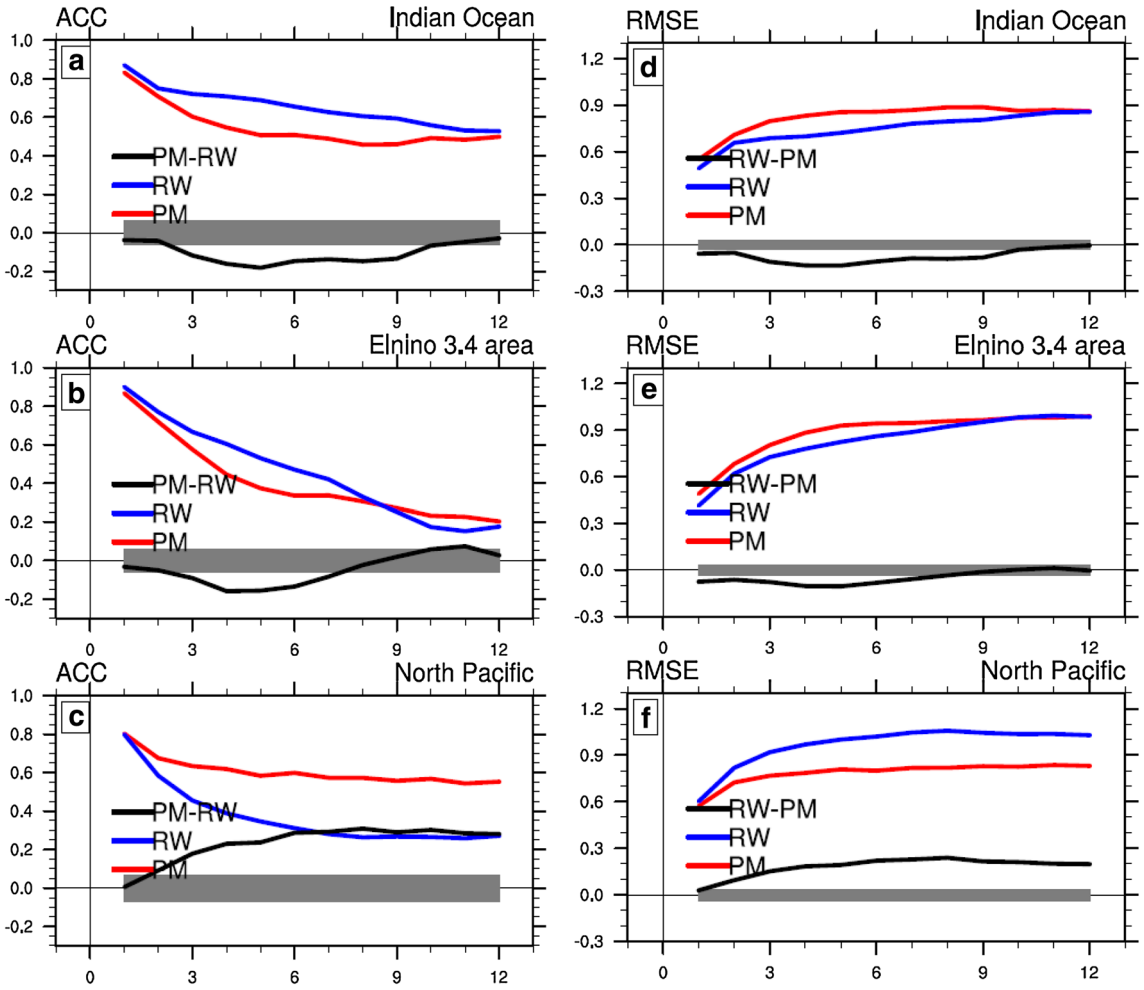

Fig. 4 ACC (left) and RMSE (right) calculated by regional averaged SSTA. The blue and red lines denote the actual skill and the perfect skill, respectively. The black lines denote the differences of forecast skill. The grey shaded bars denote the magnitude of 5% significance levels, and differences larger than the shaded bars are significant. From top to bottom: the tropical Indian Ocean (10°S–10°N, 40°E–90°E), the Nino3.4 region (5°S–5°N, 170°W–120°W), the North Pacific region (20°N–40°N, 130°E–170°W). The SST is averaged in each region first, then ACC and RMSE calculated

striking that over almost half of the grid points (more precisely about 40%), the perfect skill is lower than the actual skill (negative in y) at all the forecast times in the first 5 months. These regions are confined mainly over the tropic (red) and subtropical (blue) oceans (also see Fig. 3). Second, the skill-persistence rule holds over more than half of the grid points (the sum of the 1st and 3rd quadrants) consistently for all the lead times: that is, a higher persistence in perfect model corresponds to a higher perfect skill (1st quadrant) and a higher persistence in real world corresponds to a higher actual skill (3rd quadrant). In the meantime, there are approximately 42% of grid points in the second quadrant, inconsistent with the skill-persistence rule. It is tempting to attribute the large amount of grids in the 2nd quadrant to the effect of model bias, which tends to lower the actual skill such that a higher real world persistence could not guarantee a higher actual skill. This is, however, not necessarily the case, as will be discussed later in simple models. Third, the slopes of the regression lines in the first and third quadrants are not parallel (gray thick lines). The slope is markedly smaller in the third quadrant than in the first quadrant. This could also reflect the effect of the model bias: namely, model bias always lowers the actual skill, thus equal negative and positive persistence differences (on x coordinate) will give different skill differences (on y coordinate), less gain in skill for the actual skill than perfect skill. Finally, there are few grid points in the fourth quadrant. This implies that a higher model persistence generates a higher perfect skill with highly probability of ~90% (percentage ratio between the 1st and 4th quadrants), in contrast to the case of a higher persistence of real world, which produces a higher actual skill with a probability of ~50% (percentages ratio between

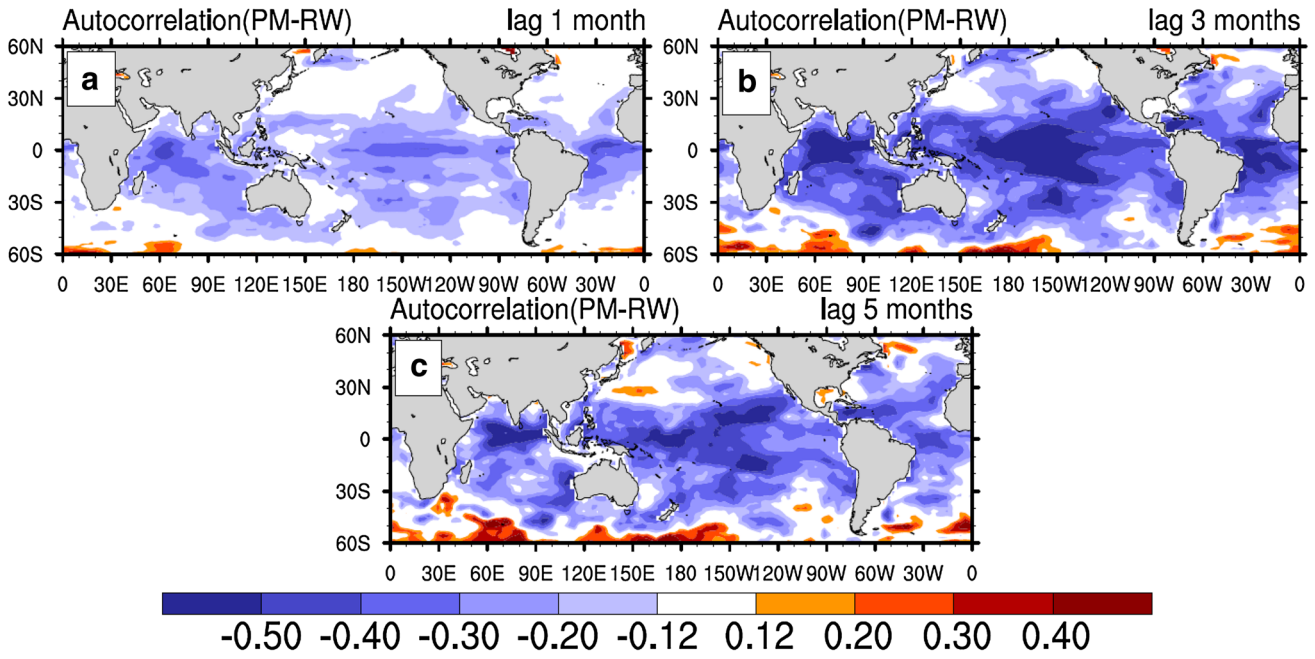

Fig. 5 Differences of SST autocorrelation between HadISST and FOAM for different lag times (PM minus RW). **a** 1 month; **b** 3 months; **c** 5 months

3rd and 2nd quadrants). This preferred high skill for perfect skill shows the skill-persistence rule holds well when the model persistence is higher than the observation, as will also be shown later in simple models.

In summary, in terms of both the ACC and RMSE, the actual skill can be higher than the perfect skill over almost half of the globe. The skill difference is related to the persistence difference, with a higher persistence in the real world favoring a higher actual skill and vice versa. However, the actual skill is suppressed systematically, presumably, partly, by the model bias. The skill-persistence rule holds well for higher model persistence (~90%), but not for higher real world persistence (~50%). Some of the major features here can be understood in terms of a simple autoregressive model in the next section.

4 Forecast skill in statistical models

We now further investigate theoretically how the forecast skill is related to the persistence. Kumar et al. (2014) interpreted the relation between the ACC of the forecast skill and persistence from a signal-to-noise ratio perspective, under many assumptions, including the same total variance of the real world and perfect model forecast. They did not investigate the relation between the forecast RMSE and persistence either. Therefore, it is unclear how robust is their conclusion in general. Here, we will show that, some major features of the relationship between forecast skill and persistence in FOAM can be understood in the AR1 framework in terms of both the ACC and RMSE.

4.1 Statistical models

We will use two simple statistical models to understand the CGCM forecasts skill by conducting analogous experiments. The models are the first order regressive model (AR1) and the second order autoregressive model (AR2) which assume the variable depends linearly on its own previous values and a stochastic noise term. The AR1 model is defined as:

$$X_{n+1} = r \cdot X_n + \varepsilon, \tag{3}$$

where X_{n+1} and X_n are model state at time $n+1$ and n, respectively, r is the model parameter that can be obtained by regressing the time series onto its previous value, and is equal to the lag-1 autocorrelation coefficient of X, ε is a white noise.

The AR2 model is defined as:

$$X_{n+1} = r_1 \cdot X_n + r_2 \cdot X_{n-1} + \varepsilon, \tag{4}$$

where the model parameters r_1 and r_2 can also be determined by the regressive method analogous to the AR1 model.

The AR1 models can be regarded as the first order approximation to the complex climate system. It has been used extensively in the understanding of the mechanism of climate variability (Hasselmann 1976) ranging from interannual variability of ENSO (e.g., Penland and Magorian 1993) to decadal variability of PDO (e.g., Newman et al. 2003).

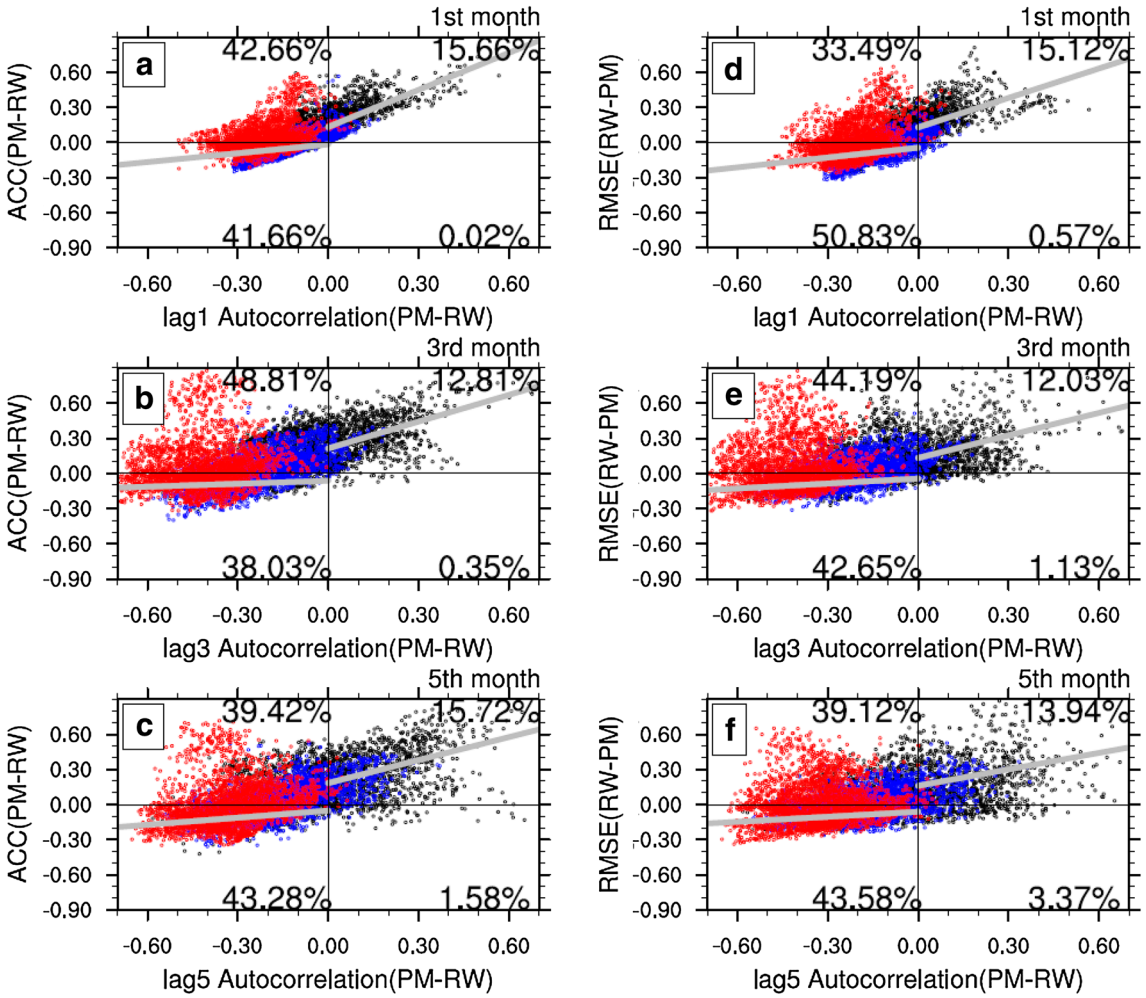

Fig. 6 Scatterplots between the differences of persistence and CGCM forecast skill. ACC and RMSE are shown in the left and right panels, respectively. The x coordinates represent the differences of autocorrelation (PM-RW) for different lag months, and the y coordinates in the left and right panel represent the differences of ACC (PM-RW) and RMSE (RW-PM) for different lead times (from top to bottom: 1 month, 3 months, 5 months), respectively. The red, blue and black dots denote the grid point between 20°S–20°N, 20°S(N)–40°S(N) and 40°S(N)–60°S(N), respectively. The number in each quadrant denotes the percentage of the grid points relative to the total grid points

4.2 Theoretical solution for AR1 model forecasts

Here we construct two AR1 models with different parameters to represent the real world and perfect model scenarios respectively. The "real world" model (superscript '*rw*') assumes the form of:

$$X_{n+1}^{rw} = r \cdot X_n^{rw} + \varepsilon_n^{rw}, \tag{5}$$

with the AR1 coefficient r being the persistence for the real world.

The "perfect model" (i.e., the forecast model, superscript '*pm*') is assumed the form of:

$$X_{n+1}^{pm} = p \cdot X_n^{pm} + \varepsilon_n^{pm}, \tag{6}$$

where the AR1 coefficient p is the persistence of the model and can be biased from the real world persistence r.

We perform the RW forecast and PM forecast as in FOAM. We first integrate the RW model (Eq. 5) and PM model (Eq. 6) once to generate the time series for the real world and the forecast model world, respectively. Then the forecast model (Eq. 6) is used to predict the real world and model world:

$$X_{n+1}^{f,rw} = p \cdot X_n^{f,rw} + \varepsilon_n^{pm}, \tag{7}$$

$$X_{n+1}^{f,pm} = p \cdot X_n^{f,pm} + \varepsilon_n^{pm}, \tag{8}$$

where the superscript "f" denotes the forecast value. These two forecasts mimic the FOAM's RW and PM forecasts, so their forecast skill can be regarded as the actual skill and the perfect skill, respectively. With infinite ensemble member

and forecast number as well as the absence of initial error, we can obtain the theoretical solutions of the ACC and RMSE at the kth step for RW forecast (see "Appendix" for more general cases) as:

$$ACC_{k,rw} = r^k, \tag{9}$$

$$(RMSE_{k,rw})^2 = 1 - 2p^k r^k + p^{2k}, \tag{10}$$

The perfect model forecast can be derived by replacing the real world persistence r with the perfect model persistence p in Eqs. 9 and 10 as:

$$ACC_{k,pm} = p^k, \tag{11}$$

$$(RMSE_{k,pm})^2 = 1 - p^{2k}, \tag{12}$$

One interesting feature in Eq. (9) is that in the AR1 framework, the ACC forecast skill for RW (Eq. 5) using the biased model (Eq. 6) completely depends on the actual persistence r and is irrelevant to the model persistence p. This occurs because, in an AR1 world, the ensemble mean forecast on each initial anomaly will decay towards zero with forecast time. If the model uses the real world as the initial condition, every initial condition decays the same rate p^k after k steps of forecast, and therefore their time correlation remains the same as their initial condition (i.e., the correlation between $X^{f,rw}_{n+k}$ and X^{rw}_{n+k} is same as the one between X^{rw}_{n} and X^{rw}_{n+k}), which is the real world persistence for actual skill here. Unlike ACC, however, the RMSE of RW forecast depends on both the real world and model persistence in Eq. (10), which can be written as:

$$(RMSE_{k,rw})^2 = (r^k - p^k)^2 + \left(1 - r^{2k}\right), \tag{13}$$

where the first term on the right hand side represents the error results from model bias, and the second term represents the impact of persistence. Equation (13) shows that a large model bias $|p–r|$ tends to induce a large RMSE, and a high real world persistence r will lead to small RMSE, in agreement with one's general expectation. Note that, since $r^k - p^k = (r-p)\cdot\sum_{n=0}^{k-1} r^n p^{k-1-n}$, for the same r and $|p–r|$, the RMSE from the model bias of $p>r$ will be larger than that of $p<r$ as the former has a larger p, suggesting a forecast model with the persistence higher than the real world tends to produce a larger forecast error than that with the persistence lower than real world.

The difference between the actual skill and perfect skill can be measured in either forecast ACC or forecast RMSE, which can be derived from Eqs. (9)–(12) as:

$$ACC_{k,rw} - ACC_{k,pm} = r^k - p^k, \tag{14a}$$

$$(RMSE_{k,pm})^2 - (RMSE_{k,rw})^2 = 2p^k\cdot(r^k - p^k). \tag{14b}$$

Equations (14a, 14b) show explicitly that the actual skill will exceed the perfect skill, equivalently in terms of ACC and RMSE, if the real world persistence is larger than the forecast model persistence ($r>p$), and vice versa. Therefore, theoretically, in an AR1 model, the skill-persistence rule holds perfectly such that the forecast skill difference can be predicted perfectly using the persistence difference.

4.3 Numerical result of AR1 model ensemble forecast

We further investigate the forecast skill in the AR1 framework, but for the practical application with a finite sample size and forecast number, analogous to the FOAM experiments. We use the AR1 analogue model to perform numerical forecast experiments as in FOAM.

We first examine the forecast skill over the three regions discussed in Fig. 4. The AR1 coefficients of the real world model and the forecast model are obtained by calculating the autocorrelation values from the year-round monthly time series of HadISST dataset and the aforementioned FOAM's "truth" (HIST run), respectively (Table 1). In each AR1 model, the "truth" and "observation" are first generated, where the "observation" is derived from the corresponding "truth" superimposed with a random "observational" error. Then, the forecasts are initialized by adding small perturbations onto the "truth" initial conditions with an ensemble size of 20. A total of 552 forecasts (46 years of 12 months' forecasts) are performed for each region as in the FOAM forecasts. Figure 7 shows that the forecast skill evolves with the forecast times. Similar to Fig. 4, the actual skill (blue lines) in the tropical Indian Ocean and Nino3.4 region are significantly higher than the perfect skill (red lines) on most forecast times and the forecast skill eventually converges. In contrast, in the North Pacific region, the difference of forecast skill is indiscernible among the actual skill and perfect skill because the persistence of the observation (0.83) and FOAM (0.79) is very close to each other.

In the theoretical AR1 world, we can also study the forecast skill with an additional experimental forecast by using the *perfect real world model* (PRWM). The skill of PRWM forecast (black lines) provides theoretically the best forecast and therefore represents the true upper bound of the actual skill. In reality, however, it is impossible to achieve

Table 1 AR1 coefficient parameters of the real world model and the forecast model

Region	Forecast model	Real world
Nino 3.4	0.738	0.948
North Pacific	0.794	0.830
Tropical Indian Ocean	0.672	0.909

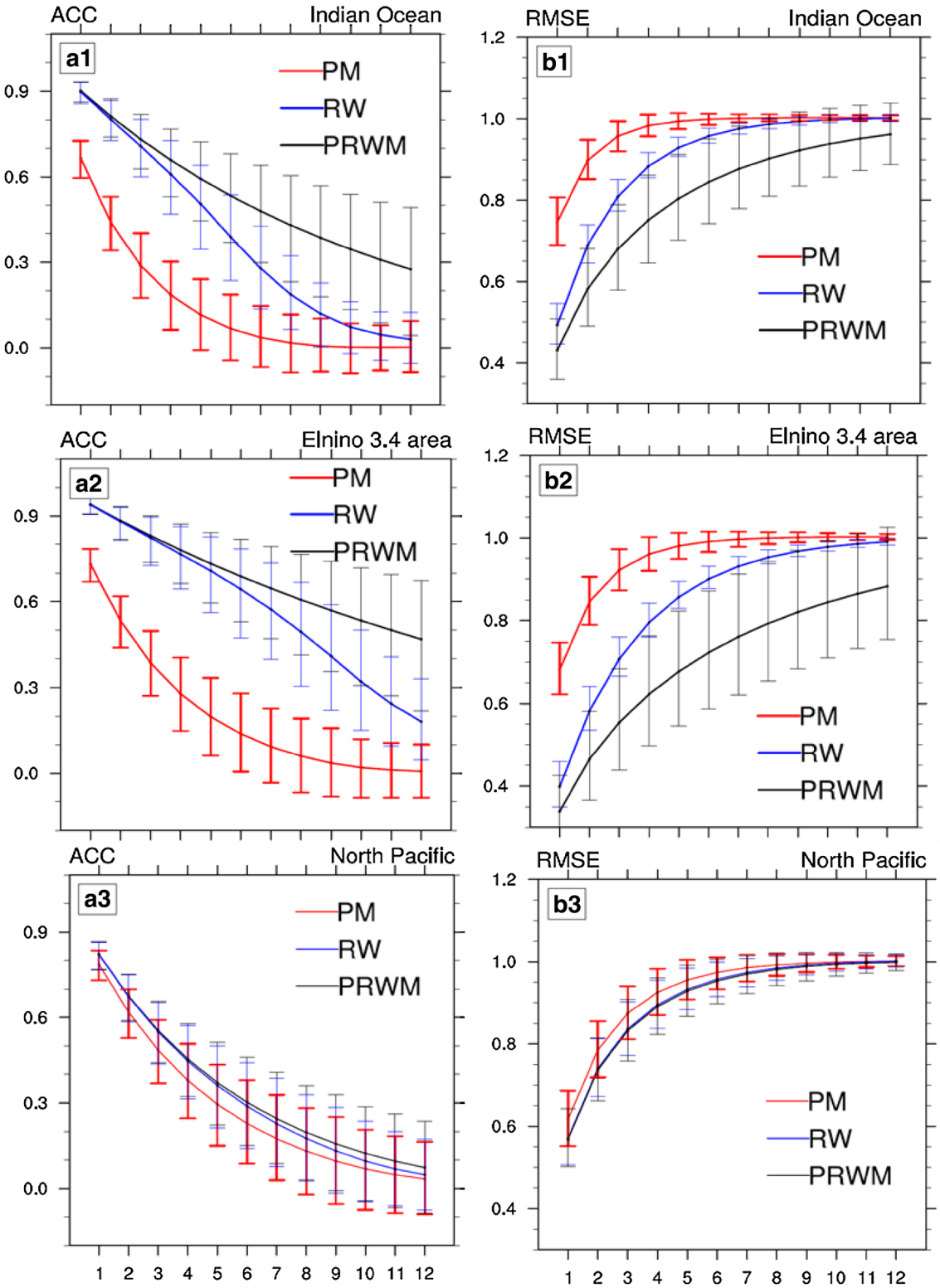

Fig. 7 ACC (left) and RMSE (right) by AR1 model forecast. The red, blue and black lines denote the PM, RW and PRWM forecast skill. From top to bottom: the tropical Indian Ocean, the Nino3.4 region, the north Pacific region. The error bars represent the 5% significance levels

the PRWM, because we can never have a perfect model for the real world. Nevertheless, it does indicate that it is critical to improve climate models for improving climate forecast in the future.

It is seen in Fig. 7 that the ACC of RW and PRWM is nearly the same at the first forecast time, consistent with the theoretical solution (with infinite ensemble size and forecast number, as well as zero initial error, the ACC of RW and PRWM will be identical at all forecast steps, see "Appendix"). However, as time increases, the skill of RW gradually deviates from PRWM and eventually converges to PM, reflects the impact of sampling error associated with finite sample size and forecast number (Kumar 2009). Note that the difference of ACCs between PM and RW in the first two regions maximizes at forecast months 5–6, which is consistent with Fig. 4 and therefore can be interpreted in the theoretical solution. Namely, for $p \approx 0.7$ and $r \approx 0.9$, $r^k - p^k$ (Eq. 14a) will reach the maximum when k equals 5–6. Also note that the ACC of RW in numerical forecast deviates from the PRWM as forecast time increases, a discrepancy from the theoretical solution that reflects the impact of sampling error due to finite ensemble size and forecast number (see "Appendix").

Now, we further apply AR1 globally to understand the FOAM experiments. We will train the AR1 model on the SST time series at each grid point and perform the ensemble RW and PM forecast experiments as in FOAM. Figure 8 shows the scatterplots for the difference of persistence and forecast skill between RW and PM forecasts as for the FOAM forecasts in Fig. 6. The AR1 model (Fig. 8) shows some resemblance to the FOAM model (Fig. 6). Overall, the skill-persistence rule holds, especially when the model persistence is larger (1st and 4th quadrants). In the AR1 model here, almost all points fall into the 1st quadrant, where a higher model persistence corresponds to a higher perfect skill, and the 3rd quadrant, where a higher RW persistence corresponds to a higher actual skill. Therefore, the skill-persistence rule holds well regardless of the difference of persistence. This is expected from the theoretical solution for AR1 model in Eqs. (14a, 14b). In FOAM, however, the scatter is much larger, especially for a higher RW persistence (quadrants 2 and 3). Some of the scatters can be seen in the AR1 model here, especially for longer lead times, because of the sampling error. A perfect skill-persistence rule would, however, constrain all points clustered around a curve through the 1st and 3rd quadrants, and this curve should be the diagonal line of slope 1 for the ACC, according to

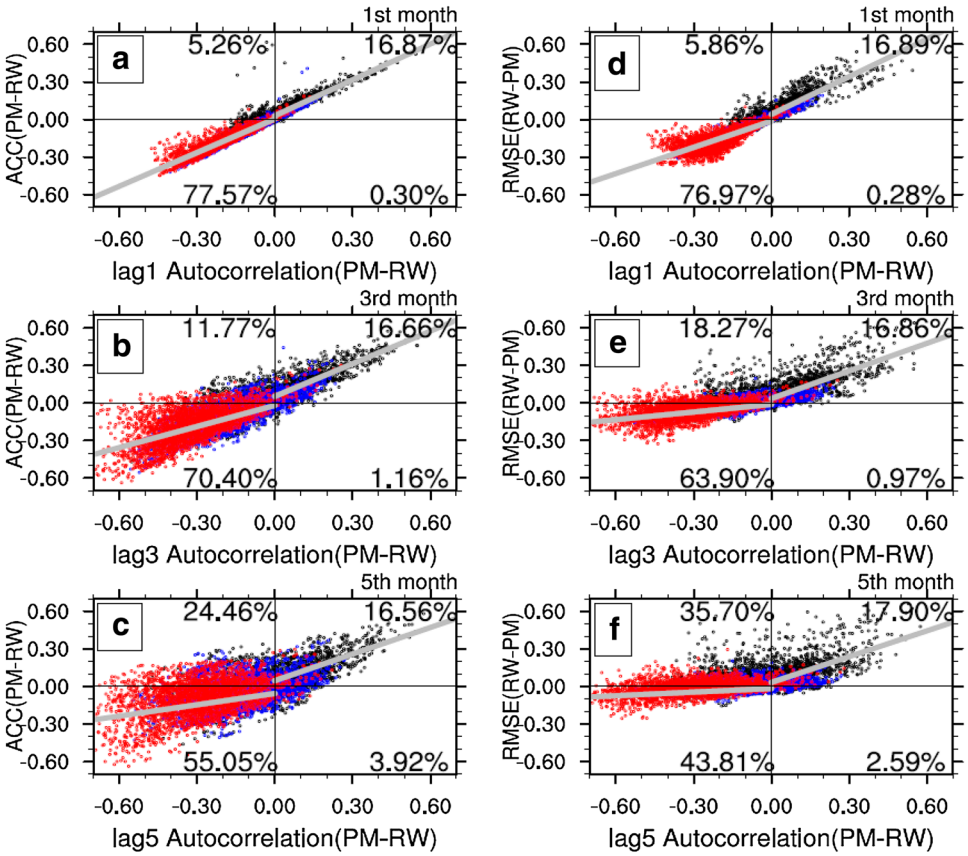

Fig. 8 Same as Fig. 6 but for AR1 model forecast

Eq. (14a), and should remain in the 1st and 3rd quadrants for RMSE according to Eq. (14b). With the finite sample size in Fig. 8, however, there are some scatters of points into the 2nd quadrant, especially at larger forecast times. Even at the first step, ~ 5% of points scatter into the 2nd quadrant (Fig. 8a, d), and the percentage of points in the 2nd quadrant increases significantly for months 3 and 5, all inconsistent with the theory of infinite sample size in Eqs. (14a, 14b). This inconsistency is contributed partly by the sampling error of the ensemble forecast, as discussed in details in the "Appendix". When the ensemble size is increased to 100, the percentages of points in the 2nd quadrant are indeed reduced in both ACC and RMSE on all forecast months (Fig. 9).

The AR1 model can explain several major features of the FOAM results. First, the AR1 model also captures the asymmetric feature of the skill-persistence slope between the 1st and 3rd quadrants in ACC (Fig. 8a–c) weaker than in RMSE (Fig. 8d–f), qualitatively similar to FOAM (Fig. 6a–c vs. d–f). This can be interpreted from Eqs. (14a and 14b). The ACC difference depends linearly on the difference of persistence $p^k - r^k$, while the RMSE difference has an additional factor of $2p^k$. Generally, for an equal $|p^k - r^k|$, the p for a positive $(p^k - r^k)$ is larger than that for a negative one, which on average leads to a larger RMSE difference for positive $(p^k - r^k)$. In FOAM, however, the model bias tends to lower both ACC and RMSE so as to cause a flatter and steeper slope in the 3rd and 1st quadrant, respectively. In addition, the percentage of RMSE in the 2nd quadrant is larger than that of ACC, while in FOAM forecast their magnitudes are very close. This reflects that RMSE are more sensitive to the sampling error than ACC in AR1 model (see "Appendix" for discussion).

The AR1 model also produces a percentage ratio between the 4th and 1st quadrants (e.g. ~ 5% for the 5th forecast step RMSE) much higher than that between the 2nd and 3rd quadrants (e.g. ~ 80% for the 5th forecast step RMSE) (Fig. 8), as in FOAM (Fig. 6). As discussed above, in AR1 model the points in the 2nd and 4th quadrants purely result from sampling error, thus it implies that the impact of sampling error for negative persistence difference ($p < r$) is stronger than the positive one ($p > r$), i.e., the sign of forecast skill difference corresponding to a positive persistence difference is less influenced by the sampling error. This can be understood from the sampling error as discussed in more details in the "Appendix".

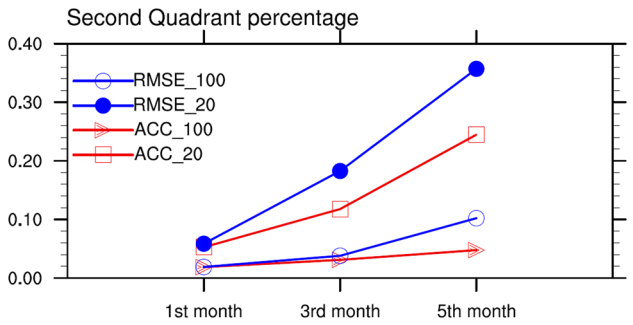

Fig. 9 The percentages of the second quadrants, varies with different ensemble size (from 20 to 100) and forecast times

In spite of its success, the AR1 model also has important difference from the FOAM results. The most striking difference between FOAM and the AR1 model is much less points in the 2nd quadrant in the latter, especially in the ACC, e.g., with less than 10% in AR1 compared with ~ 40% in FOAM at the 1st month. This suggests that the large number of points in the 2nd quadrant in FOAM is caused by the complexity of the FOAM model beyond the description of AR1. As a simple test of the importance of processes beyond AR1 for the FOAM experiments, we repeat all the AR1 model forecasts with the AR2 model forecasts, as shown in Fig. 10. It can be seen that the percentage of the points in the 2nd quadrant is increased significantly from AR1 to AR2 forecasts in both ACC and RMSE, especially at larger forecast times. Moreover, the slopes of ACC and RMSE in the 3rd quadrant are much flatter in the AR2 model, resembling the FOAM results more than the AR1 model.

In summary, in the AR1 framework, the skill-persistence rule holds perfectly well, allowing some scattering due to sampling error. The application of AR1 to FOAM forecast suggest that the AR1 model, and the skill-persistence rule, can explain the overall feature of persistence-forecast skill relation in FOAM. However, the large number of points in the 2nd quadrant in FOAM cannot be explained by the AR1 model, implying the importance of other processes.

The skill-persistence rule also holds for low order autoregressive moving average model (ARMA(1,1)). The ARMA(1,1) model is defined as:

$$X_{n+1} = \alpha \cdot X_n + \varepsilon_n + \beta \cdot \varepsilon_{n-1}, \tag{15}$$

where α and β are model parameters, ε_n and ε_{n-1} are white noise at step n and $n-1$, respectively. Thus the real world model can be written as:

$$X^{rw}_{n+1} = r \cdot X^{rw}_n + \varepsilon^{rw}_n + \beta_r \cdot \varepsilon^{rw}_{n-1}, \tag{16}$$

and the forecast model:

$$X^{pm}_{n+1} = p \cdot X^{pm}_n + \varepsilon^{pm}_n + \beta_p \cdot \varepsilon^{pm}_{n-1}. \tag{17}$$

With assumption of infinite ensemble size and forecast number, as well as zero initial error, the analytical solutions of ACC and RMSE for ARMA(1,1) model can be readily obtained:

$$ACC_{k,rw} - ACC_{k,pm} = \rho_{rw,k} - \rho_{pm,k}, \tag{18a}$$

$$(RMSE_{k,pm})^2 - (RMSE_{k,rw})^2 = 2p^k \cdot (\rho_{rw,k} - \rho_{pm,k}), \tag{18b}$$

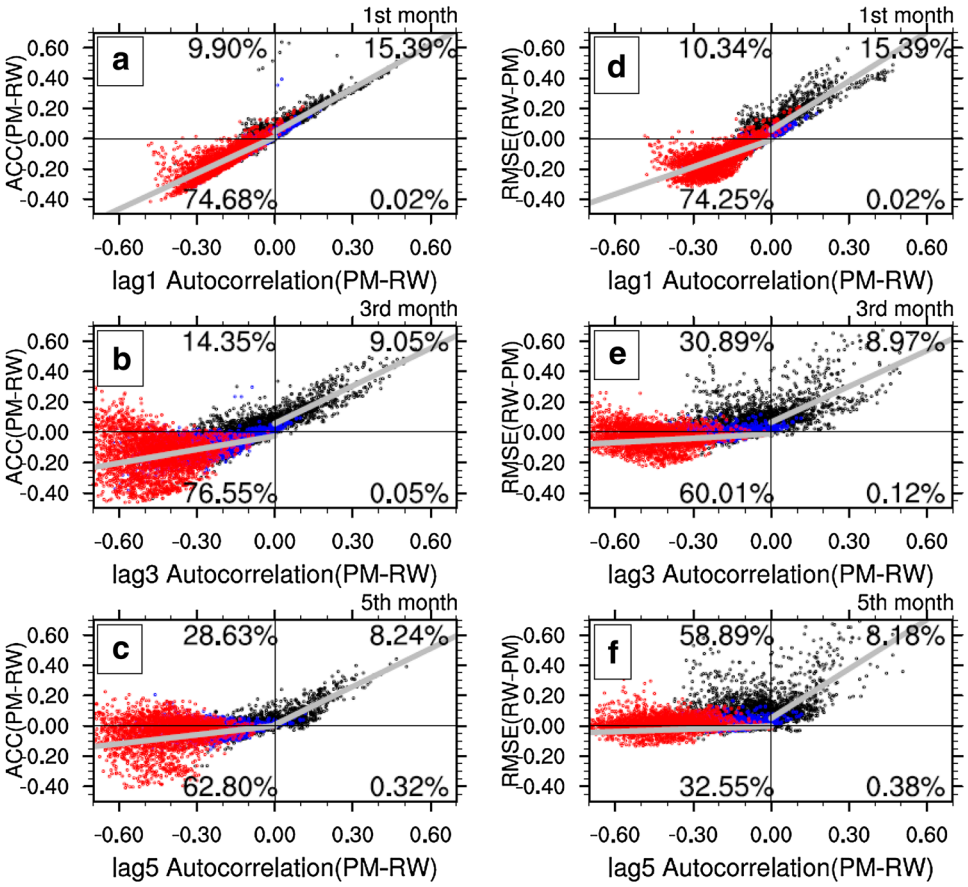

Fig. 10 Same as Fig. 8, but for AR2 model

where $\rho_{rw,k}$ and $\rho_{pm,k}$ are the autocorrelation of the real world model and forecast model at lag k, respectively, which can be written as:

$$\rho_{rw,k} = r^k + r^{k-1}\beta_r \frac{1-r^2}{1+\beta_r{}^2+2r\beta_r}, \tag{19a}$$

$$\rho_{pm,k} = p^k + p^{k-1}\beta_p \frac{1-p^2}{1+\beta_p{}^2+2p\beta_p}. \tag{19b}$$

Therefore, the skill-persistence relationship in ARMA(1,1) model is similar to that of AR1 (Eqs. 14a and 14b). Namely, the difference of ACC between RW and PM forecasts in ARMA(1,1) model is identical to the difference of real world persistence and forecast model persistence, and the difference of RMSE at forecast step k equals to the difference of persistence multiplied by $2p^k$.

5 Summary and discussion

In this paper, the factors contributing to the difference between the actual skill and perfect skill are investigated in fully CGCM seasonal prediction experiments. The difference in forecast skill, which are measured in both ACC and RSME, are understood in terms of its relationship with the difference of persistence. The relationship between the persistence and forecast skill is further understood in the framework of a simple AR1 model. The major conclusions are as follows.

1. For seasonal forecasts, there are large areas where the actual skill exceeds the perfect skill, such as the tropical Pacific and Indian Ocean (in FOAM). Indeed, the area of higher actual skill can be comparable with the area of higher perfect skill. Therefore, the perfect skill may not serve as an accurate indicator of "room for improvement", as suggested by Kumar et al. (2014).
2. For seasonal prediction, the difference of forecast skill and persistence is related through a skill-persistence rule. This rule states that, a higher actual persistence corresponds to a higher actual forecast skill, and vice versa. Therefore, the regions of higher actual skill are usually caused by a higher persistence in the real world than in the model, and vice versa.
3. The skill-persistence rule tends to be distorted in more complex models such as a CGCM, in particular, for the case of higher actual persistence. There are too many points in the 2nd quadrant in FOAM such that a higher

real world persistence is still far from a guarantee of a higher actual skill (with about 50% probability). This excessive points in the 2nd quadrant is caused partly by the sampling error and partly by more complex processes beyond AR1. These complex model bias and model processes tend to suppress the actual skill.

As mentioned above, the perfect skill in this study is different from the traditional approach (e.g., Kumar et al. 2014). By using the AR1 model, we further derive the analytical solutions of ACC and RMSE for traditional approach (see 'Appendix"). It can be found that for traditional approach, the difference of actual skill and perfect skill depends not only on the persistence of real world and forecast model, but also on the total variance of real world and perfect model forecast. When the total variance of real world and perfect model forecast are equal, which is a key assumption of Kumar et al. (2014), the perfect skill of our approach is identical to that of traditional approach. Moreover, if the variance of forecast model equals or is larger than the real world variance, the skill-persistence rule holds for the traditional approach, namely a higher SST persistence in the real world than in model could produce a higher actual skill. This again suggests that there is not necessarily a relation between the actual skill and perfect skill. The perfect skill is model dependent and can be an erroneous estimate of the true upper limit of real world prediction skill. The "perfect model" and "perfect skill" may not provide a reliable upper limit of real skill unless the model precisely captures the major statistical properties, including at least the persistence, as in the observation.

Further issues remain to be studied. The most important question is why there are so many points in the 2nd quadrant in FOAM model and what processes beyond AR1 is responsible for this distortion of the relation between forecast skill and persistence. The non-AR1 processes may also allow a transient growth of initial error and thus impact the forecast skill with lead time.

Acknowledgements This work is supported by the National Basic Research Program of China (2017YFA0603801), the National Key R&D Program of China (2016YFE0102400), the Special Fund for Public Welfare Industry (GYHY201506012), the Basic Research Fund of CAMS (2015Z002) and US NSF Climate Dynamics 1656907.

Appendix: derivation of perfect skill and actual skill in AR1 model

In this section, we will derive the perfect and actual skill in the case of infinite forecasts number. We first derive the forecast ACC and RMSE for RW forecast in the presence of initial error and sampling error. Recalling Eqs. (6) and (8), assume the forecast starts from step n of the truth of real world, i.e., X_n^{rw}, and now we want to forecast the state of step $n+k$. Then, the truth at step $n+k$ is X_{n+k}^{rw}, and the ensemble mean of the forecast is $X_{n+k}^{f,\,rw}$. Assuming the ensemble mean of initial condition error is $\varepsilon_{i,n}^{rw}$, then we have:

$$X_{n+k}^{rw} = r^k X_n^{rw} + \sum_{j=0}^{k-1} (r^{k-j-1}\varepsilon_{n+j}^{rw}), \tag{20}$$

$$\overline{X_{n+k}^{f,\,rw}} = p^k X_n^{rw} + \frac{1}{m}\sum_{j=0}^{k-1}\left(p^{k-j-1}\sum_{l=1}^{m}\varepsilon_{l,n+j}^{pm}\right) + p^k\varepsilon_{i,n}^{rw}, \tag{21}$$

where m is the ensemble size, $\varepsilon_{l,n+j}^{pm}$ is the noise of ensemble member l at step $n+j$. Then the variance of $\overline{X_{n+k}^{f,\,rw}}$ is:

$$\begin{aligned}
\left\langle \overline{X_{n+k}^{f,\,rw}}, \overline{X_{n+k}^{f,\,rw}} \right\rangle &= \left\langle p^k X_n^{rw} + \frac{1}{m}\sum_{j=0}^{k-1}\left(p^{k-j-1}\sum_{l=1}^{m}\varepsilon_{l,n+j}^{pm}\right) + p^k\varepsilon_{i,n}^{rw},\ p^k X_n^{rw} + \frac{1}{m}\sum_{j=0}^{k-1}\left(p^{k-j-1}\sum_{l=1}^{m}\varepsilon_{l,n+j}^{pm}\right) + p^k\varepsilon_{i,n}^{rw} \right\rangle \\
&= p^{2k}\left\langle X_n^{rw}, X_n^{rw}\right\rangle + \left\langle \frac{1}{m}\sum_{j=0}^{k-1}\left(p^{k-j-1}\sum_{l=1}^{m}\varepsilon_{l,n+j}^{pm}\right), \frac{1}{m}\sum_{j=0}^{k-1}\left(p^{k-j-1}\sum_{l=1}^{m}\varepsilon_{l,n+j}^{pm}\right)\right\rangle + p^{2k} \\
&\quad \times \left\langle \varepsilon_{i,n}^{rw}, \varepsilon_{i,n}^{rw}\right\rangle = p^{2k}\left\langle X_n^{rw}, X_n^{rw}\right\rangle + \frac{1}{m^2}\sum_{j=0}^{k-1}\left(p^{2(k-j-1)}\sum_{l=1}^{m}\left\langle \varepsilon_{l,n+j}^{pm}, \varepsilon_{l,n+j}^{pm}\right\rangle\right) + p^{2k}\left\langle \varepsilon_{i,n}^{rw}, \varepsilon_{i,n}^{rw}\right\rangle \\
&= p^{2k}\left\langle X_n^{rw}, X_n^{rw}\right\rangle + \frac{1}{m}\sum_{j=0}^{k-1} p^{2(k-j-1)}\langle \varepsilon^{pm}, \varepsilon^{pm}\rangle + p^{2k}\left\langle \varepsilon_{i,n}^{rw}, \varepsilon_{i,n}^{rw}\right\rangle = p^{2k}\left\langle X_n^{rw}, X_n^{rw}\right\rangle \\
&\quad + \frac{1}{m}\left(1-p^2\right)\left(\sum_{j=0}^{k-1} p^{2(k-j-1)}\right)\left\langle X_n^{pm}, X_n^{pm}\right\rangle + p^{2k}\left\langle \varepsilon_{i,n}^{rw}, \varepsilon_{i,n}^{rw}\right\rangle = \left[p^{2k}\left(1+{c_{rw}}^2\right) + \frac{d}{m}\left(1-p^{2k}\right)\right]\left\langle X_n^{rw}, X_n^{rw}\right\rangle,
\end{aligned} \tag{22}$$

where '<>' denotes the covariance upon infinite size of forecasts, and $d = \langle X_n^{pm}, X_n^{pm}\rangle / \langle X_n^{rw}, X_n^{rw}\rangle$ is the ratio between the variance of forecast model and real world X, ${c_{rw}}^2 = \langle \varepsilon_{i,n}^{rw},\ \varepsilon_{i,n}^{rw}\rangle / \langle X_n^{rw},\ X_n^{rw}\rangle$ is the ratio between the variance of initial error and real world X. Note that we assume that the state X_n^{rw} is independent of the noise and initial error, also the noise and initial error are uncorrelated.

Then, the ACC of RW forecast at forecast step k is:

$$ACC_{rw,k} = \frac{\left\langle \overline{X_{n+k}^{f,\ rw}}, X_{n+k}^{rw} \right\rangle}{\sqrt{\left\langle \overline{X_{n+k}^{f,\ rw}}, \overline{X_{n+k}^{f,\ rw}} \right\rangle \cdot \left\langle X_{n+k}^{rw},\ X_{n+k}^{rw} \right\rangle}} = \frac{\left\langle p^k X_n^{rw} + \frac{1}{m}\sum_{j=0}^{k-1}\left(p^{k-j-1}\sum_{l=1}^{m}\varepsilon_{l,n+j}^{pm}\right) + p^k\varepsilon_{i,n}^{rw}, r^k X_n^{rw} + \sum_{j=0}^{k-1}\left(r^{k-j-1}\varepsilon_{n+j}^{rw}\right)\right\rangle}{\sqrt{\left\langle \overline{X_{n+k}^{f,\ rw}}, \overline{X_{n+k}^{f,\ rw}} \right\rangle \cdot \left\langle X_{n+k}^{rw},\ X_{n+k}^{rw} \right\rangle}}$$
$$= \frac{p^k r^k \left\langle X_n^{rw}, X_n^{rw}\right\rangle}{\sqrt{\left\langle \overline{X_{n+k}^{f,\ rw}}, \overline{X_{n+k}^{f,\ rw}} \right\rangle \cdot \left\langle X_{n+k}^{rw},\ X_{n+k}^{rw} \right\rangle}} = \frac{r^k}{\sqrt{1 + {c_{rw}}^2 + \frac{d}{m}\left(\frac{1}{p^{2k}} - 1\right)}}, \tag{23}$$

and the RMSE at forecast step k is:

$${RMSE_{rw,k}}^2 = \frac{\left\langle \overline{X_{n+k}^{f,\ rw}} - X_{n+k}^{rw}, \overline{X_{n+k}^{f,\ rw}} - X_{n+k}^{rw} \right\rangle}{\left\langle X_{n+k}^{rw},\ X_{n+k}^{rw} \right\rangle}$$
$$= \left(p^k - r^k\right)^2 + \left(1 - r^{2k}\right) + \frac{d}{m}\left(1 - p^{2k}\right) + p^{2k}{c_{rw}}^2. \tag{24}$$

Similarly, the ACC and RMSE of PM forecast at step k:

$$ACC_{pm,k} = \frac{\left\langle \overline{X_{n+k}^{f,\ pm}}, X_{n+k}^{pm} \right\rangle}{\sqrt{\left\langle \overline{X_{n+k}^{f,\ pm}}, \overline{X_{n+k}^{f,\ pm}} \right\rangle \cdot \left\langle X_{n+k}^{pm},\ X_{n+k}^{pm} \right\rangle}}$$
$$= \frac{p^k}{\sqrt{1 + {c_{pm}}^2 + \frac{1}{m}\left(\frac{1}{p^{2k}} - 1\right)}}, \tag{25}$$

$${RMSE_{pm,k}}^2 = \frac{\left\langle \overline{X_{n+k}^{f,\ pm}} - X_{n+k}^{pm}, \overline{X_{n+k}^{f,\ pm}} - X_{n+k}^{pm} \right\rangle}{\left\langle X_{n+k}^{pm},\ X_{n+k}^{pm} \right\rangle}$$
$$= \left(1 - p^{2k}\right) + \frac{1}{m}\left(1 - p^{2k}\right) + p^{2k}{c_{pm}}^2, \tag{26}$$

where ${c_{pm}}^2 = \left\langle \varepsilon_{i,n}^{pm},\ \varepsilon_{i,n}^{pm}\right\rangle / \left\langle X_n^{pm},\ X_n^{pm}\right\rangle$ is the ratio between the variance of initial error and X in perfect mode forecast.

With the assumption of perfect initial condition and infinite ensemble size, we derive Eqs. (9)–(13) as:

$$ACC_{rw,k} = r^k, \tag{27}$$

$$ACC_{pm,k} = p^k, \tag{28}$$

$$RMSE_{rw,k}^2 = \left(p^k - r^k\right)^2 + \left(1 - r^{2k}\right), \tag{29}$$

$$RMSE_{pm,k}^2 = \left(1 - p^{2k}\right). \tag{30}$$

Now, we consider an additional case, namely the *perfect real world model* case for the prediction of the truth of the real world. In this case the ACC and RMSE can be derived by simply replacing p of Eqs. (28) and (30) by r:

$$ACC_{prwm,k} = r^k, \tag{31}$$

$${RMSE_{prwm,k}}^2 = 1 - r^{2k}, \tag{32}$$

where '*prwm*' denotes forecast by *perfect real world model* (PRWM). We can see that the ACC of PRWM forecast is identical to RW forecast, however, the difference of RMSE is:

$${RMSE_{rw,k}}^2 - {RMSE_{prwm,k}}^2 = \left(p^k - r^k\right)^2 \geqslant 0, \tag{33}$$

This indicates that it is the forecast skill by PRWM (prediction in the perfect model as the real world), instead of PM (prediction in a biased model), provides the upper bound of actual skill. Therefore, a model provides the true upper bound for actual skill only if the model can produce the correct statistical property, at least, the persistence, as the real world.

Now we discuss the impact of sampling error on forecast skill, and explain why the percentage of RMSE in the 2nd quadrant is larger than that of ACC, as well as why the percentage ratio between the 4th and 1st quadrants are higher than that between the 2nd and 3rd quadrants. To simplify the discussion, we assume the initial error is zero. From Eqs. (23) and (25), the difference of ACC between RW and PM forecast can be written as:

$$ACC_{rw,k} - ACC_{pm,k} = \frac{r^k}{\sqrt{1+\frac{d}{m}\left(\frac{1}{p^{2k}}-1\right)}} - \frac{p^k}{\sqrt{1+\frac{1}{m}\left(\frac{1}{p^{2k}}-1\right)}} = \frac{1}{\sqrt{1+\frac{1}{m}\left(\frac{1}{p^{2k}}-1\right)}}\left(\frac{r^k}{a}-p^k\right), \tag{34}$$

where

$$a = \sqrt{1+\frac{(d-1)\left(\frac{1}{p^{2k}}-1\right)}{m+\left(\frac{1}{p^{2k}}-1\right)}}. \tag{35}$$

If $r>p$, the sign of $ACC_{rw,k} - ACC_{pm,k}$ depends on the sign of $(d-1)$. When $d<1$, we have $a<1$ such that $\frac{r^k}{a} - p^k > 0$ and $ACC_{rw,k} - ACC_{pm,k} > 0$, indicating that the points could not move from the 3rd quadrant (infinite sampling size) to the 2nd quadrant (finite sampling size). However, when $d>1$, which is usually the case in FOAM, $a<1$, then $\frac{r^k}{a} - p^k$ might be smaller than zero, thus the points could shift into the 2nd quadrant.

In the case of $r<p$ and $d>1$, we have $a>1$ and, in turn, $\frac{r^k}{a} - p^k < 0$. Thus, points could not move from the 1st quadrant into the 4th quadrant. When $d<1$, $\frac{r^k}{a} - p^k$ might be larger than zero, however, since this will rarely happen in FOAM, the percentage of 4th quadrant will be small.

The difference of RMSE is:

$$RMSE_{pw,k}{}^2 - RMSE_{rw,k}{}^2 = 2p^k\left(r^k - p^k\right) + \frac{1-d}{m}\left(1-p^{2k}\right). \tag{36}$$

Since in FOAM for a large number of points $d>1$, if $r>p$, $RMSE_{pw,k}{}^2 - RMSE_{rw,k}{}^2$ could be negative, then the points might move from the 3rd quadrant to the 2nd quadrant. When $r<p$ and $d>1$, $RMSE_{pw,k}{}^2 - RMSE_{rw,k}{}^2 < 0$, the points will be constrained in the 1st quadrant. If $d<1$, $RMSE_{pw,k}{}^2 - RMSE_{rw,k}{}^2$ could be smaller than zero, as discussed above, this will seldom occur.

Note that the sign of the RMSE difference is more sensitive to the sampling error than ACC, especially for small ensemble size and large forecast step k. This could be seen from Eqs. (34) and (36). For large k, the first tern in the righthand side of Eqs. (36) is a small term, while the second term approximates to $\frac{1-d}{m}$, thus the sign of $(1-d)$ directly determines the sign of the RMSE difference. However, for ACC, the sign of ACC difference depends not only on d, but also on r^k and p^k. Consider an extreme case of $k\to\infty$ and $r>p$, then $a\to\sqrt{d}$, the sign of $\frac{r^k}{a} - p^k$ reverses that of $r^k - p^k$ only when $\sqrt{d} > r^k/p^k$.

Now we derive the ACC and RMSE by traditional perfect model approach (e.g., Kumar et al. 2014). For simplicity we discuss the case of infinite ensemble size with zero initial error.

Take one member of the RW ensemble as "truth":

$$X^{f,rw}_{n+k} = p^k X^{rw}_n + \sum_{j=0}^{k-1}\left(p^{k-j-1}\varepsilon^{pm}_{n+j}\right). \tag{37}$$

Recall Eq. (21), with assumption of infinite ensemble size and zero initial error, the ensemble mean of RW forecasts is:

$$\overline{X^{f,\,rw}_{n+k}} = p^k X^{rw}_n. \tag{38}$$

Then we have:

$$\left\langle X^{f,rw}_{n+k}, \overline{X^{f,\,rw}_{n+k}}\right\rangle = p^{2k}\left\langle X^{rw}_n, X^{rw}_n\right\rangle, \tag{39}$$

$$\left\langle X^{f,rw}_{n+k}, X^{f,rw}_{n+k}\right\rangle = p^{2k}\left\langle X^{rw}_n, X^{rw}_n\right\rangle + \left(1-p^{2k}\right)\left\langle X^{pm}_n, X^{pm}_n\right\rangle = \left[(1-d)p^{2k}+d\right]\left\langle X^{rw}_n, X^{rw}_n\right\rangle, \tag{40}$$

$$\left\langle \overline{X^{f,\,rw}_{n+k}}, \overline{X^{f,\,rw}_{n+k}}\right\rangle = p^{2k}\left\langle X^{rw}_n, X^{rw}_n\right\rangle, \tag{41}$$

$$\left\langle X^{f,rw}_{n+k} - \overline{X^{f,\,rw}_{n+k}}, X^{f,rw}_{n+k} - \overline{X^{f,\,rw}_{n+k}}\right\rangle = \left(1-p^{2k}\right)\left\langle X^{pm}_n, X^{pm}_n\right\rangle = d\left(1-p^{2k}\right)\left\langle X^{rw}_n, X^{rw}_n\right\rangle, \tag{42}$$

where $d = \left\langle X^{pm}_n, X^{pm}_n\right\rangle / \left\langle X^{rw}_n, X^{rw}_n\right\rangle$.

Then the ACC and RMSE of PM forecasts by traditional approach can be written as:

$$ACC^t_{pm,k} = \frac{\left\langle \overline{X^{f,\,rw}_{n+k}}, X^{f,rw}_{n+k}\right\rangle}{\sqrt{\left\langle \overline{X^{f,\,rw}_{n+k}}, \overline{X^{f,\,rw}_{n+k}}\right\rangle, \left\langle X^{f,rw}_{n+k}, X^{f,rw}_{n+k}\right\rangle}} = \frac{p^{2k}\left\langle X^{rw}_n, X^{rw}_n\right\rangle}{p^k\sqrt{(1-d)p^{2k}+d}\left\langle X^{rw}_n, X^{rw}_n\right\rangle} = \frac{p^k}{\sqrt{(1-d)p^{2k}+d}}, \tag{43}$$

$$\left(RMSE^t_{pm,k}\right)^2 = \frac{\left\langle X^{f,rw}_{n+k} - \overline{X^{f,\,rw}_{n+k}}, X^{f,rw}_{n+k} - \overline{X^{f,\,rw}_{n+k}}\right\rangle}{\left\langle X^{f,rw}_{n+k}, X^{f,rw}_{n+k}\right\rangle} = \frac{d\left(1-p^{2k}\right)\left\langle X^{rw}_n, X^{rw}_n\right\rangle}{\left[(1-d)p^{2k}+d\right]\left\langle X^{rw}_n, X^{rw}_n\right\rangle} = 1 - \frac{p^{2k}}{(1-d)p^{2k}+d}. \tag{44}$$

The difference of ACC and RMSE between traditional approach and our method for the perfect model forecast can therefore be written as:

$$\left(ACC^t_{pm,k}\right)^2 - \left(ACC_{pm,k}\right)^2 = \frac{(1-d)\left(p^{2k}-p^{4k}\right)}{(1-d)p^{2k}+d}, \tag{45}$$

$$\left(RMSE^{t}_{pm,k}\right)^2 - \left(RMSE_{pm,k}\right)^2 = \frac{(d-1)\left(p^{2k}-p^{4k}\right)}{(1-d)p^{2k}+d}, \quad (46)$$

and the difference of ACC and RMSE between perfect model forecast and real world forecast by traditional approach are:

$$\left(ACC^{t}_{pm,k}\right)^2 - \left(ACC_{rw,k}\right)^2 = \left(p^{2k}-r^{2k}\right) + \frac{(1-d)\left(p^{2k}-p^{4k}\right)}{(1-d)p^{2k}+d}, \quad (47)$$

$$\left(RMSE^{t}_{pm,k}\right)^2 - \left(RMSE_{rw,k}\right)^2 = 2p^{k}\left(r^{k}-p^{k}\right) + \frac{(d-1)\left(p^{2k}-p^{4k}\right)}{(1-d)p^{2k}+d}. \quad (48)$$

When $d=1$, i.e. the same total variance in the real world and perfect model, the ACC and RMSE of traditional approach (Eqs. 43 and 44) are identical to our approach (Eqs. 28 and 30). If $d>1$ ($d<1$), i.e. the variance of the perfect model forecast is larger (smaller) than real world, the perfect skill of traditional approach is lower (higher) than that of our approach (Eqs. 45 and 46). Thus, for the traditional approach, the difference between the perfect skill and actual skill depends on not only on p and r, but also on the variances of the real world and perfect model. If $d \geqslant 1$ and $p<r$, $(ACC^{t}_{pm,k})^2 - (ACC_{rw,k})^2 < 0$ and $(RMSE^{t}_{pm,k})^2 - (RMSE_{rw,k})^2 > 0$, the skill-persistence rule holds for the traditional approach.

References

Anderson JL (2001) An ensemble adjustment Kalman filter for data assimilation. Mon Weather Rev 129:2884–2903

Anderson JL (2003) A local least squares framework for ensemble filtering. Mon Weather Rev 131:634–642

Becker EJ, Dool HVD, Peña M (2013) Short-term climate extremes: prediction skill and predictability. J Clim 26:512–531

Becker E, Dool HVD, Zhang Q (2014) Predictability and forecast skill in NMME. J Clim 27:5891–5906

Boer G, Kharin VV, Merryfield WJ (2013) Decadal predictability and forecast skill. Clim Dyn 41:1817–1833

Chen M, Wang W, Kumar A (2010) Prediction of monthly-mean temperature: the roles of atmospheric and land initial conditions and sea surface temperature. J Clim 23:717–725

Dunstone NJ, Smith DM (2010) Impact of atmosphere and sub-surface ocean data on decadal climate prediction. Geophys Res Lett 37:L02709. https://doi.org/10.1029/2009GL041609

Gaspari G, Cohn SE (1999) Construction of correlation functions in two and three dimensions. Q J R Meteorol Soc 125:723–757

Griffies S, Bryan K (1997) A predictability study of simulated North Atlantic multidecadal variability. Clim Dyn 13:459–487

Hasselmann K (1976) Stochastic climate models. Part I: theory. Tellus 28:473–485

Holland MM, Blanchard-Wrigglesworth E, Kay J, Vavrus S (2013) Initial-value predictability of Antarctic sea ice in the Community Climate System Model 3. Geophys Res Lett 40:2121–2124. https://doi.org/10.1002/grl.50410.f

Jacob R (1997) Low frequency variability in a simulated atmosphere ocean system. Ph.D. dissertation, University of Wisconsin–Madison, p 155

Kumar A (2009) Finite samples and uncertainty estimates for skill measures for seasonal prediction. Mon Weather Rev 137:2622–2631

Kumar A, Peng P, Chen M (2014) Is There a relationship between potential and actual skill? Mon Wea Rev 142:2220–2227

Liu Z, Kutzbach J, Wu L (2000) Modeling climate shift of El Niño variability in the Holocene. Geophys Res Lett 27:2265–2268

Liu Z, Otto-Bliesner B, Kutzbach J, Li L, Shields C (2003) Coupled climate simulations of the evolution of global monsoons in the Holocene. J Clim 16:2472–2490

Liu Z, Liu Y, Wu L, Jacob R (2007) Seasonal and long-term atmospheric responses to reemerging North Pacific Ocean variability: a combined dynamical and statistical assessment. J Clim 20:955–980

Liu Y, Liu Z, Zhang S, Rong X, Jacob R, Wu S, Lu F (2014) Ensemble-based parameter estimation in a coupled GCMusing the adaptive spatial average method. J Clim 27:4002–4014

Lu F, Liu Z, Liu Y, Zhang S, Jacob R (2016) Understanding the control of extratropical atmospheric variability on ENSO using a coupled data assimilation approach. Clim Dyn. https://doi.org/10.1007/s00382-016-3256-7

Mehta VM, Suarez MJ, Manganello JV, Delworth TL (2000) Oceanic influence on the North Atlantic Oscillation and associated Northern Hemisphere climate variations: 1959–1993. Geophys Res Lett 27:121–124

Meinshausen M, Smith S et al (2011) The RCP GHG concentrations and their extension from 1765 to 2300. Clim Change. https://doi.org/10.1007/s10584-011-0156-z

Newman M, Compo GP, Alexander MA (2003) ENSO-forced variability of the Pacific decadal oscillation. J Clim 16:3853–3857

Pegion K, Sardeshmukh PD (2011) Prospects for improving subseasonal predictions. Mon Weather Rev 139(11):3648–3666

Penland C, Magorian T (1993) Prediction of Nino-3 sea surface temperature using linear inverse modeling. J Clim 6:1067–1076

Pohlmann H, Kröger J, Greatbatch RJ, Müller WA (2016) Initialization shock in decadal hindcasts due to errors in wind stress over the tropical Pacific. Clim Dyn 49:2685–2693

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geophys Res 108(D14):1063–1082. https://doi.org/10.1029/2002JD002670

Sévellec F, Fedorov AV (2013) Model bias reduction and the limits of oceanic decadal predictability: importance of the deep ocean. J Clim 26:3688–3707

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93:485–498

Teng H, Branstator G, Meehl GA (2011) Predictability of the Atlantic overturning circulation and associated surface patterns in two CCSM3 climate change ensemble experiments. J Clim 24:6054–6076

Tobis M, Schafer C, Foster I, Jacob R, Anderson J (1997) FOAM: expanding the horizons of climate modeling. Supercomputing 1997 conference, Supercomputing, ACM/IEEE 1997 Conference, pp 27–27

Wu L, Liu Z, Gallimore R, Jacob R, Lee D, Zhong Y (2003) Pacific decadal variability: the tropical mode and the North Pacific mode. J Clim 16:1101–1120

Younas W, Tang Y (2013) PNA predictability at various time scales. J Clim 26:9090–9114. https://doi.org/10.1175/JCLI-D-12-00609.1

Zhang S, Harrison MJ, Rosati A, Wittenberg A (2007) System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. Mon Weather Rev 135:3541–3564