Grand European and Asian-Pacific multi-model seasonal forecasts: maximization of skill and of potential economical value to end-users

Andrea Alessandri[1,2], Matteo De Felice[2], Franco Catalano[2], June-Yi Lee[3,4], Bin Wang[3], Doo Young Lee[5], Jin-Ho Yoo[6], and Antije Weisheimer[7]

[1]Royal Netherlands Meteorological Institute, De Bilt, The Netherlands
[2]Agenzia Nazionale per le nuove Tecnologie, l'energia e lo sviluppo economico sostenibile, Roma, Italy
[3]International Pacific Research Center, Honolulu, HI, USA
[4]Institute of Environmental Studies, Pusan National University, Busan, South Korea
[5]Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC-CNS), Barcelona, Spain
[6]Asian-Pacific Economic Cooperation Climate Center (APCC), Busan, South Korea
[7]European Center For Medium Range Weather Forecasts, Shinfield, England, United Kingdom

Manuscript submitted to

Climate Dynamics

May 23, 2017

Corresponding author:
Andrea Alessandri (KNMI, ENEA)
P.O. Box 201
NL-3730 AE, De Bilt, Netherlands
E-mail: andrea.alessandri@knmi.nl
Phone: +31 30 2206 685
Fax: +31 30 2210 407

Abstract

Multi-Model Ensembles (MMEs) are powerful tools in dynamical climate prediction as they account for the overconfidence and the uncertainties related to single-model ensembles. Previous works suggested that the potential benefit that can be expected by using a MME amplifies with the increase of the independence of the contributing Seasonal Prediction Systems. In this work we combine the two MME Seasonal Prediction Systems (SPSs) independently developed by the European (ENSEMBLES) and by the Asian-Pacific (APCC/CliPAS) communities. To this aim, all the possible multi-model combinations obtained by putting together the 5 models from ENSEMBLES and the 11 models from APCC/CliPAS have been evaluated.

The grand ENSEMBLES-APCC/CliPAS MME enhances significantly the skill in predicting 2m temperature and precipitation compared to previous estimates from the contributing MMEs. Our results show that, in general, the better combinations of SPSs are obtained by mixing ENSEMBLES and APCC/CliPAS models and that only a limited number of SPSs is required to obtain the maximum performance. The number and selection of models that perform better is usually different depending on the region/phenomenon under consideration so that all models are useful in some cases. It is shown that the incremental performance contribution tends to be higher when adding one model from ENSEMBLES to APCC/CliPAS MMEs and vice versa, confirming that the benefit of using MMEs amplifies with the increase of the independence the contributing models.

To verify the above results for a real world application, the Grand ENSEMBLES-APCC/CliPAS MME is used to predict retrospective energy demand over Italy as provided by TERNA (Italian Transmission System Operator) for the period 1990-2007. The results demonstrate the useful application of MME seasonal predictions for energy demand forecasting over Italy. It is shown a significant enhancement of the potential economic value of forecasting energy demand when using the better combinations from the Grand MME by comparison to the maximum value obtained from the better combinations of each of the two contributing MMEs. The above results demonstrate for the first time the potential of the Grand MME to significantly contribute in obtaining useful predictions at the seasonal time-scale.

1 Introduction

Two well-validated Multi-Model Ensemble (MME) Seasonal Prediction Systems have been independently compiled by the European and by the Asian-Pacific communities. The Climate Prediction and its Application to Society (CliPAS) project sponsored by the Asian-Pacific Economic Cooperation Climate Center (APCC) assembled a well-validated MME prediction system by putting together 14 independent modeling tools from the Asian-Pacific community (Wang et al., 2009). The five "State-of-the-Art" coupled models developed in Europe in the framework of the European Commission FP7 project ENSEMBLES composed the second MME (Weisheimer et al., 2009; Alessandri et al., 2011b). Much of the prediction systems participating in CliPAS and in ENSEMBLES have also joined the multi-model operational efforts that are being established to provide real-time seasonal forecasts by the EUROpean Seasonal to Inter-annual Prediction (EUROSIP; Vitart et al., 2007), the North American MultiModel Ensemble (NMME; Kirtman et al., 2014) and the APCC operational seasonal MME (APCC MME; Min et al., 2017). Such operational development follows from the recognition that the MME strategy is indeed a viable approach for improving performance in ENSO prediction (Jin et al., 2008) and for adequately resolve forecast uncertainty (Kirtman et al., 2014). Previous works showed that MMEs are powerful tools in dynamical climate prediction (Palmer et al., 2004; Weisheimer et al., 2009) and that they are more effective in enhancing performance with the increase of the independence of the contributing Seasonal Prediction Systems (SPSs; Wang et al., 2009; Alessandri et al., 2011b). The multi-models get their performance from the skill of the contributing models, so that MME skill is generally proportional to the mean skill of the individual models (Yoo and Kang, 2005). However, the relation between single-model averages and MME skill is not linear and the multi-model performance is superior to the average of the Single-Model Ensembles (SMEs). As explained in Hagedorn et al. (2005a), this is mainly attributable to error cancellations and to the nonlinearity of the skill metrics applied. The independence of the contributing models between each other is a prerequisite to obtain error cancellations (Hagedorn et al., 2005a) and for skill amplification to occur (Yoo and Kang, 2005). Weigel et al. (2008) showed that multi-models act by gradually widening the ensemble spread (i.e. reducing the over-confidence of the single models) and moving the ensemble mean toward truth without reducing the potential predictability. The increasing diversity of SPSs performance contributes to the higher predictive skills of the MME (Alessandri et al., 2011b). Even though all models are based on similar approximations of the same dynamical equations, a considerable source of errors in seasonal forecast arises from uncertainties due to model formu-

lation, from uncertainties due parameterizations of unresolved sub-grid scale processes and from uncertainties due to model initialization. The idea behind the MME is that if the uncertainties of the models are independent from each other, the associated model errors may be random in nature; thus, a MME approach may cancel out the errors contained in individual models. In this respect, independence is here intended as a synonym of diversity in models formulation, parameterization, and initialization.

In order to be useful for decision-making, seasonal climate predictions need to be probabilistic and the capability of probability forecasts to provide valuable information to end-users needs to be assessed (e.g: Richardson, 2006). At the decision-making level, probability forecasts are regarded by virtue of their potential economic value. This notion of value is conceptually different from the notion of skill in the meteorological sense. In fact, the potential economic value cannot be assessed by analyzing meteorological variables alone; it depends also on the users economic parameters.

By joining together the two independently developed MME Seasonal Prediction Systems, this work aims at maximising the prediction performance currently attainable to obtain robust climate services. Section 2 defines the methodology by first describing the grand ENSEMBLES-APCC/CliPAS multi-model (sub-section 2.1). Then, the evaluation method and the observations/reanalysis data used as reference are introduced in subsection 2.2. The results of this paper are discussed in Section 3. Section 3.1 compares the deterministic performances between the ENSEMBLES vs. APCC/CliPAS MMEs and evaluate the skill improvement in the grand ENSEMBLES-APCC/CliPAS MME. The maximization of the probabilistic performance using the Grand MME predictions is reported in Section 3.2, which also discusses independence and the incremental contributions of the single models to the probabilistic performance (subsection 3.2.1). Section 3.3 evaluates the advantages in terms of Potential Economic Value (PEV) of using the Grand MME in forecasting the energy load over Italy. Finally, in Section 4 a summary of the main conclusions is given.

2 Method

2.1 The grand ENSEMBLES-APCC/CliPAS multi-model

The one-tier hindcasts from the Asian Pacific (APCC/CliPAS, Wang et al., 2009; Min et al., 2014; Lee et al., 2015), and European (ENSEMBLES, Weisheimer et al., 2009; Alessandri et al., 2011b) communities has been collected into a grand MME covering the common 1983-2005 hindcast period. The contributing institutions and the references of the 11 Seasonal Prediction Systems (SPS)

coming from APCC/CliPAS and of the five SPSs from ENSEMBLES are summarized in Table 1. The Grand MME combinations are obtained using equal-weights for each of the SPSs coming from ENSEMBLES and from APCC/CliPAS. For APCC/CliPAS, in addition to the retrospective forecasts from the seven one-tier SPSs already described in Wang et al. (2009), the hindcasts from four further systems have been added from member organisations that subsequently contributed to the operational APCC MME (Min et al., 2014; Lee et al., 2015). The additional coupled systems, compared to Wang et al. (2009), come from South-Korean centers of Pusan National University (PNU, Ahn and Kim, 2014) and APCC (Jeong et al., 2008, 2012) and from the Canadian Meteorological service (Merryfield et al., 2013). All institutions provide one seasonal prediction system, except for the Meteorological Service of Canada, which provides two systems based on the version 3 (CanCM3) and the version 4 (CanCM4) of the Canadian Atmospheric General Circulation Model (AGCM), respectively.

Retrospective forecasts, starting 1 May and 1 November are used and the variables of interest, i.e. 2 meter temperature and precipitation, are interpolated to common latitude-longitude regular grid ($2.5^\circ \times 2.5^\circ$). For each forecasting system, Table 2 reports the original horizontal resolution and the number of ensemble members provided. As shown in Table 2, the MMEs considered in this work are formed by a variable number of ensemble members, ranging from 5 to 30. All available members made available from each center are used when computing ensemble means to be used in assessing correlation skill. The number of ensemble members provided by a SPS may affect its probabilistic skill (see Palmer et al., 2004). To avoid this effect, the analysis on the probabilistic skill have been carried out considering nine members for all SPSs, with the exception of SNU and PNU (in Winter) that have a smaller size (Table 2). For the SPSs that provided more members, the 9-member ensembles considered in the analysis were sampled by taking members 1-9. We also checked the effect on the results when the 9-members are sampled randomly, instead of getting first 9 members, finding no appreciable effects on the results of this paper (not shown).

2.2 Evaluation method and Data

The performance of 1-month lead seasonal forecasts is evaluated by taking the ECMWF ERA-INTERIM reanalysis (Berrisford et al., 2009) as the reference for 2m temperature and the Global Precipitation Climatology Project (GPCP; Adler et al., 2003) satellite-based observations for precipitation. The deterministic skill is measured by computing the correlations of the ensemble-mean forecasts with the reference data. The probabilistic performance of the grand ENSEMBLES-APCC/CliPAS MME is measured by the Brier Skill Score (Wilks, 2011). For a more accurate

estimation of the probabilistic performances, a leave-one-out cross-validation procedure has been implemented, excluding each target year from the computation of terciles and climatological mean of the sample distribution. We also compared the results when not using the leave-one-out technique and quite interestingly we found no appreciable overestimation of the scores (not shown). All the calculations are performed using the forecast anomalies, computed for each contributing model by removing the corresponding climatology from the original ensemble forecasts. A similar process is applied to the verification reanalysis/observation data.

To assess the potential usefulness of probabilistic forecasts, we applied the PEV metric by assuming a cost/loss model related to a binary event, as described in Richardson (2003). Electricity demand data used here have been provided by TERNA (Italian TSO, Transmission System Operator) and they refer to the period 1990-2007. The hourly data provided by TERNA are subdivided in eight regions over Italy: North-West, North, North-East, North-Center, Center, South, Sicily and Sardinia. The data have been aggregated over the Italian domain and, as we are focusing on summer demand, the monthly demand has been calculated summing up all the hourly-loads for each month. Only June and July have been retained, while August has not been included due to industrial closure. In fact during August industrial facilities usually close in Italy for one or two weeks reducing electricity demand independently of temperature. Given that during 1990-2007 the electricity demand was steadily increasing, a trend removal has been performed by fitting a second-order regression model for each region and then computing the deviation from the fit (i.e. regression residuals) as described in De Felice et al. (2015).

3 Results

3.1 ENSEMBLES vs. APCC/CliPAS: independence and summation of deterministic performance

The ENSEMBLES and APCC/CliPAS MMEs are compared and the gain of using the grand ENSEMBLES-APCC/CliPAS MME in terms of deterministic performance is assessed in this section. Here, the results for surface air temperature in boreal summer (June-July-August; hereinafter JJA) are reported. Note that the results for JJA are also well representative of the performance of the winter season, therefore in the following only JJA is reported for brevity.

Figure 1a shows the correlation skill of the ENSEMBLES MME for 1-month lead JJA seasonal forecasts of 2m temperature. The results show that the skill tends to be concentrated over tropical Pacific, and from there tends to irradiate toward the whole tropical belt and extratropics. Other than

Pacific, good correlation skill is found over northern Indian Ocean, Northern China and Mongolia, northwestern Atlantic and Euro-Mediterranean basin. The largest performance tends, however, to be confined over the ocean. The skill difference between ENSEMBLES and APCC/CliPAS and its frequency distribution are reported in Figure 1b and Figure 1c, respectively. Overall, the global scale performance of APCC/CliPAS and ENSEMBLES are comparable (Figure 1c) but they show regional differences (Figure 1b). The skill difference between APCC/CliPAS and ENSEMBLES displays quite a patchy pattern, with positive differences in some regions, which tend to be compensated by negative values in other areas. For instance, APCC/CliPAS displays increased correlation over northern Eurasia, some parts of North and South Atlantic and equatorial eastern Indian oceans. On the other hand ENSEMBLES tends to be better over the Euro-Mediterranean, northeastern China, northern Indian subcontinent and northwestern tropical Pacific. The regional differences between APCC/CliPAS and ENSEMBLES MMEs give evidence of the independence of the models contributing to the two MMEs. As discussed in previous works, the independence of the contributing models between each other is a prerequisite to obtain skill amplification and error cancellations in MMEs (Hagedorn et al., 2005b; Alessandri et al., 2011b). Therefore it is expected an increase of the performance by collecting APCC/CliPAS and ENSEMBLES into a grand-MME.

Figure 2 shows the skill difference between the grand ENSEMBLES-APCC/CliPAS and ENSEMBLES MMEs. The grand MME improves in much of the areas even if in some regions ENSEMBLES still appears to perform slightly better (Figure 2a). Overall, ENSEMBLES-APCC/CliPAS is seen to improve skill in the majority of the grid points as shown by the distribution of the point-by-point differences (Figure 2b). The improvement in the grand ENSEMBLES-APCC/CliPAS MME is found both on all (sea and land; red) and on land only (light blue) grid points as indicated by the vertical dashed lines in Figure 2b, representing the average value of the differences distribution. As summarized in Table 3, the averaged improvements are similar with respect to both APCC/CliPAS and ENSEMBLES. Overall, considering the average of sea and land grid-points, the Grand MME improves by 5% compared to APCC/CliPAS and ENSEMBLES in both DJF and JJA. The improvement tends to be higher over land-only grid points: it is 6% in DJF and 8% in JJA against APCC/CliPAS, while the land improvement, compared with ENSEMBLES, amounts to 5% in DJF and 7% in JJA.

3.1.1 Optimal combination of the ENSEMBLES and APCC/CliPAS systems for ensemble-mean predictions.

Combining the two MME Seasonal Prediction Systems independently developed by the European and by the Asian-Pacific communities makes it possible to assess the maximum level of skill that is currently attainable for seasonal predictions. To this aim, all the possible MME combinations ($\sum_{k=1,16} C_{16}^{k} = 65535$) have been evaluated by putting together the 5 models from ENSEMBLES and the 11 models from APCC/CliPAS and using equal-weights for each SPS in the Grand MME. Figure 3 shows the Pattern Correlation Coefficients (PCCs) computed over the (a) Northern Middle Latitude (NML; 25N-75N), (b) Tropics (25S-25N) and (c) Southern Middle Latitude (SML; 75S-25S) regions as a function of the number of models and obtained with all the possible combinations of the models available. Blue filled-circles represent the cases in which the combinations are obtained with only models from ENSEMBLES, while red filled-circles are for the combinations of the APCC/CliPAS models only. The combinations mixing models from both APCC/CliPAS and ENSEMBLES are the green filled circles. For both temperature (Figure 3) and precipitation (Figure 4), the combinations from the grand ENSEMBLES-APCC/CliPAS MME that mix models from ENSEMBLES and APCC/CliPAS provide the larger values of PCCs. The maximum performance obtained by mixing ENSEMBLES and APCC/CliPAS models (green dashed lines) considerably improves what would be obtained by ENSEMBLES only (blue dashed line) or by APCC/CliPAS only (red dashed lines) in all domains. It should be noted here that there might be a higher chance of getting higher scores simply as a consequence of the larger number of model combinations in the Grand MME. An ad-hoc significance test has been implemented to verify that the improvement is indeed related to mixing models from independent sources. The null hypothesis of getting as high or higher improvements as a consequence of the larger number of model combinations has been tested through a monte carlo method by re-shuffling the models and choosing randomly the 5-models group (synthetic ENSEMBLES MME) and the 11-models group (synthetic APCC/CliPAS MME) from the Grand 16-models pool (1000 repetitions). The 5^{th} and 95^{th} percentiles of the null hypothesis distribution are reported in Table 5 together with the actual improvement obtained through the Grand MME (i.e. Maximum Grand MME minus Maximum ENSEMBLES or APCC/CliPAS performances). It is shown that, with the only exception of precipitation over Southern Middle Latitudes, the null hypothesis can be rejected at the 5% significance level, therefore indicating that the improvement of the Grand MME is at least in part a consequence of mixing models from the two diverse groups. For all regions, the performance

initially increases by adding models to the ensembles, while tending to level off when passing a given threshold of models after which the skill decreases for the combinations of more models. The optimal number of models required to maximise performance is different depending on the region and variable under consideration. The optimal combination of ENSEMBLES-APCC/CliPAS models to forecast temperature scores a maximum PCC of 0.359 in the NML, 0.520 in the Tropics and 0.423 in the SML. This improves significantly (5% significance level) over what was achievable by using ENSEMBLES only (0.313 for NML, 0.492 for Tropics and 0.382 for SML) and APCC/CliPAS only (0.327 for NML, 0.496 for Tropics and 0.352 for SML). For the Tropics, the maximum performance tends to level off when adding more than 4 models and with the best combination composed by 5 models obtained by mixing ENSEMBLES and APCC/CliPAS. Then, the skill decreases steadily for the combinations of more than 6 models. Middle latitudes display a larger number of models required to maximise PCC for temperature and with the best combination composed by 8 models over both NML and SML. For precipitation, the mix of 7 ENSEMBLES and APCC/CliPAS models is required to obtain the maximum skill over Tropics. On the other hand, over NML the maximum PCC performance for precipitation is obtained with only 5 models, while 9 models are required for SML.

It is noteworthy that the performance obtained by simply combining all models in ENSEMBLES, APCC/CliPAS, or the Grand MMEs (diamonds) are considerably lower in all domains compared to the optimal combinations of the grand MME models. The democratic inclusion of all models in ENSEMBLES and APCC/CliPAS is also outperformed by the optimal combination of a limited number of models in each group, with the only exception of ENSEMBLES in the SML domain. In fact, the ensemble average of all ENSEMBLES systems also coincides with the maximum performance considering all the combinations of the five ENSEMBLES models over SML (Figs. 3c and 4c). Together with the fact that the ENSEMBLES MME is composed by a smaller number of models, this appears to be related to the smaller spread in the performance of the ENSEMBLES models compared with APCC/CliPAS (Figs. 3c and 4c). It follows that none of the models from ENSEMBLES behaves as negative outliers such that to degrade the performance of the ensemble combinations.

Several works showed that the ability of dynamical models in simulating Sea Surface Temperature (SST) mean state over tropical regions can impact the skill in predicting the interannual anomalies of temperature and precipitation (Lee et al., 2010; Alessandri et al., 2011a). To evaluate the relationship between climatological SST bias over Tropics and the performance of the MME combinations, the PCC for temperature (Figure 5) and precipitation (Figure 6) are displayed for

all the MME combinations as a function of the respective SST bias over the Tropics. The linear fit of all the MME combinations is indicated when significance of the slope of linear relationship is verified at the 5% level using a Fisher parametric test. Indeed, fitted lines in Figure 5 and 6 show a clear relationship between performance and tropical SST bias over both Tropics itself (panel b) as well as over NML (panel a) and SML (panel c) regions. The only exception is for the ENSEMBLES MME over NML, where the relationship between SST bias and precipitation skill is not statistically significant at 5% level (Figure 6a). The present analysis confirms that realistic mean state of SST over the Tropics is a key aspect for the models to better simulate/predict interannual tropical climate variability as well as related teleconnections over middle latitudes.

3.2 Maximization of the probabilistic performance using the Grand MME predictions

The ENSEMBLES and APCC/CliPAS are compared and the gain of using the grand ENSEMBLES-APCC/CliPAS MME in terms of overall probabilistic accuracy is evaluated. As discussed in Section 2.1, this analysis has been performed considering only nine ensemble members for each SPS to exclude the influence of ensemble size on the probabilistic skills. Here, the results for above-normal (i.e., above upper tercile of the sample distribution) 2m temperature and precipitation in boreal summer (June-July-August; hereinafter JJA) are reported. Note that the results for above-normal conditions (i.e., above upper tercile of the sample distribution) are similar to those for below-normal cases (not shown). Similarly, the results for JJA are well representative of the performance of the winter season, therefore in the following only JJA is reported for brevity.

Figure 7a shows the maximum Brier Skill Score (BSS) of 1-month lead 2m temperature seasonal forecasts started May 1st that is achievable using the grand ENSEMBLES-APCC/CliPAS MME. The BSS is obtained for each grid point by getting the maximum value of all the possible 65535 combinations using the 5 models from ENSEMBLES and the 11 models from APCC/CliPAS. The results show that the probabilistic forecast skill tends to be concentrated over tropical Pacific quite similarly to the deterministic score (see Section 3.1). This confirms that much of the skill of present dynamical seasonal climate forecasts comes from their ability in predicting ENSO (e.g. Alessandri et al., 2011b; Lee et al., 2011). The high skill tends to irradiate toward the whole tropical belt and to the extratropics from the ENSO region with largest performances located over oceans. The BSS percent gain obtained using the grand ENSEMBLES-APCC/CliPAS MME by comparison with the maximum performance of the contributing ENSEMBLES or APCC/CliPAS MMEs (whichever is larger at each grid point) is reported in Figure 7b for JJA. A considerable improvement of the grand MME is shown, with broad enhancements exceeding 20% over mid to high latitudes. It

is found a remarkable BSS increases over land areas including China, Middle East, Europe and northern North America in JJA (Figure 7b), and including Africa, Europe, Asian boreal forests, and mid-latitude North America in DJF (Not shown). As summarised in Table 5 (upper rows), the area averaged improvement of BSS for 2m temperature in boreal summer (winter) is 8.4% (7.0%) over Tropics, 10.6% (8.6%) over NML and 10.5% (10.3%) over SML.

The maximum BSS for precipitation of the 1-month lead seasonal forecasts started May 1st that is achievable using the grand ENSEMBLES-APCC/CliPAS MME is reported in Figure 8a. It is shown that, also for precipitation, the BSS maximized using the grand MME is positive almost everywhere, therefore being better than climatological non-informative forecasts. Consistently with the results for 2m temperature, the better BSS for precipitation is concentrated over equatorial Pacific and from there it irradiates toward the whole tropical belt and to some extent also to the extra-tropics. A considerable improvement of the grand MME is shown, with BSS increases exceeding 30% over land in Europe, Middle East and South East Asia in JJA (Figure 8b), and in Africa, Europe and Asian boreal forests in DJF (Not shown). Table 5 (lower rows) further reports the averaged improvements in forecasting precipitation for boreal summer (winter): the averaged % improvement amounts to 9.8 (10.4) over Tropics, 10.8 (11.5) over NML and 10.1 (10.9) over SML.

3.2.1 Incremental contributions of the single models to probabilistic performance

To characterize the contribution of each model to probabilistic performance, we compute the incremental changes in BSS of including a given model in each combination where it is not already present (sometimes reported in the literature as marginal changes, e.g. Wilks, 2011). The averaged percent-incremental change of adding a given model (j) to all 32767 possible combinations ($\mathbf{A}_{Combs}$; eqs. 1-2) where it is not already included (hereinafter incremental contribution of a model; $\Delta\overline{m}_j$ in eq. 1) is reported in Figure 9 and Figure 10 for temperature and precipitation, respectively.

$$\Delta\overline{m}_j = \frac{1}{|\mathbf{A}_{Combs}|} \sum_{i \in \mathbf{A}_{Combs}} \Delta m_{i,j} \tag{1}$$

$$\mathbf{A}_{Combs} = \left\{ \binom{|\mathbf{S}_G - \{j\}|}{k}, \forall k = 1, 2, ..., N_G - 1 \; in \; \mathbf{S}_G - \{j\} \right\} \tag{2}$$

where $\Delta m_{i,j}$ is the marginal improvement of adding the model (j) to a given combination (i) of models, $\binom{|\mathbf{S}_G - \{j\}|}{k}$ denotes the set of all k-combinations of $\mathbf{S}_G - \{j\}$, $N_G = 16$ the total number of elements in the Grand ENSEMBLES-APCC/CliPAS set of models ($\mathbf{S}_G$), and the vertical bars

|| on each side denotes the number of elements (cardinality) of a set.

In all cases, either considering NML (Fig. 9a and 10a), Tropics (Fig. 9b and 10b) or SML (Fig. 9c and 10c), the majority of models appear to add performance to the MMEs. Remarkably, all models are useful in improving the performance and appear to provide added skill at least for some regions and variables. On the other hand, a limited number of models may perform not well in some cases and we may want to remove them from the ensembles when degrading the multi-model performance. The fraction of the global grid points for which each model is necessary to maximise the performance is reported in Figure 11. Indeed, even if some models are better than others, all contribute to the maximisation of the skill in more than 10% of the grid points for both temperature (Figure 11a) and precipitation (Figure 11b). Note that we restricted the analysis only to the grid points where the skill is sufficiently good after the maximization is performed; a BSS threshold of 0.3 is considered in Figure 11 and the effect of changing the threshold to 0.1, 0.2, and 0.4 have been as well checked finding no appreciable changes in the outcomes of this analysis (not shown).

To evaluate the relative independence between the two MMEs, we compare the relative effect of including each model into the combinations of SPSs either solely from the same or solely from the other MME. Specifically, the incremental skill of adding models from one MME to all possible combinations composed exclusively by models from the other MME are compared with the respective incremental skill of adding to combinations of models solely from the same MME. We therefore compute the normalized incremental contribution of adding APCC/CliPAS or ENSEMBLES SPSs (*added* in eq. 3) to combinations of APCC-only, ENSEMBLES-only and mixed ENSEMBLES-APCC/CliPAS MMEs (*target* in eq. 3):

$$\Delta\overline{m}^{target}_{added} = \frac{1}{|\mathbf{A}^{target}_{Combs}| \times N_{added}} \sum_{j=1,N_{added}} \sum_{i \in \mathbf{A}^{target}_{Combs}} \frac{\Delta m_{i,j}}{\Delta\overline{m}_j} \tag{3}$$

$$\mathbf{A}^{target}_{Combs} = \left\{ \binom{\mathbf{S}_{target} - \{j\}}{k}, \forall k = 1, 2, ..., N^*_{target} \, in \, \mathbf{S}_{target} - \{j\} \right\} \tag{4}$$

where *target* can either refer to APCC/CliPAS (A), ENSEMBLES (E) or Grand MME (G) combinations; *added* stands for either APCC/CliPAS (A) or ENSEMBLES (E) and with $\mathbf{S}_{target}$ indicating the sets of all SPSs composing the APCC/CliPAS ($\mathbf{S}_A$; with $N_A = 11$), the ENSEMBLES ($\mathbf{S}_E$; with $N_E = 5$) or the Grand ENSEMBLES-APCC/CliPAS ($\mathbf{S}_G$; with $N_G = 16$) MMEs. Note that $N^*_{E,A,G} = N_{E,A,G}$ for $\Delta\overline{m}^E_A$ and $\Delta\overline{m}^A_E$ while $N^*_{E,A,G} = N_{E,A,G} - 1$ in

all other cases. Figure 12 displays that the normalized incremental skill contribution are significantly (10% significance level) larger when adding (red) APCC/CliPAS or (green) ENSEMBLES to target combinations of (left) ENSEMBLES-only, (right) APCC-only and (center) mixed ENSEMBLES-APCC/CliPAS. The adimensional ratios have at the denominator the total number of combinations for which either APCC/CliPAS or ENSEMBLES significantly prevail in adding skill for each domain. It is clearly shown that the incremental contribution of adding independent SPSs to the MMEs leads to significantly larger improvements of the skill in most cases over both Tropics (Fig. 12b) and middle latitudes (NML, Fig. 12a; SML, Fig. 12c). Adding ENSEMBLES (APCC/CliPAS) SPSs to the combinations of APCC-only (ENSEMBLES-only) models improves significantly (10% significance level) more than adding APCC/CliPAS (ENSEMBLES) in 86% (95%), 89% (96%), 63% (92%) of the times over NML, Tropics and SML, respectively. The above results are consistent with the idea that by mixing independent SPSs from European and Asian-Pacific communities can lead to considerable more chance of skill amplification compared to adding more models from the same community. Also consistently, the marginal contributions to the mixed ENSEMBLES-APCC/CliPAS combinations tends to be balanced, with only APCC/CliPAS performing slightly better, in particular over SML. This is probably related to the larger number of models in APCC/CliPAS, which increases the chance to effectively introduce incremental contributions to skill from the single-models. It is out of the scope of this paper to evaluate the diversity of the models in APCC/CliPAS with respect to ENSEMBLES in their full complexity. However, from Table 2 it appears clear how APCC/CliPAS and ENSEMBLES MMEs may add diversity to each other both in terms of model formulation/parameterization and in terms of initialisation. In particular, it is shown that European SPSs tend to use ECMWF reanalyses to start the hindcasts, whereas Asian-Pacific models use NCEP reanalyses in most cases. To check that the results in Figure 12 is not just an artifact due to the different size in the two MMEs, we repeated the analysis with only 5 models in each MME. By selecting randomly 5 models from the APCC/CliPAS pool (1000 repetitions), we found consistent results and no appreciable effect on the outcomes of this analysis (not shown).

3.3 Potential Economic Value of electricity load forecasts using the grand ENSEMBLES-APCC/CliPAS MME

Given the necessity of ensuring the balance between electricity production and demand, an accurate estimation of future seasonal-mean climate state can improve the efficiency and reliability of energy management at local and national scales. In fact, climate is a crucial factor in deter-

mining both the generation and demand of electricity (Rothstein and Halbig, 2010; Dubus, 2010). In Section 3.2, it was reported that there is a good probabilistic skill of the Grand ENSEMBLES-APCC/CliPAS MME over Euro-Mediterranean basin. The effectiveness of the Grand ENSEMBLES-APCC/CliPAS MME seasonal predictions in enhancing electricity demand forecasts is here assessed in terms of the Potential Economic Value (PEV; Richardson, 2006) to end users. The PEV measures the economic saving the user can make using the forecasts when faced to a binary (yes/no) event. Here we use the relative savings with respect to having only the climatological information (PEV=0) and with the maximum attainable savings that would result from perfect deterministic forecasts (PEV=1; for further details see Richardson, 2006). The two binary events of getting above normal (i.e. above upper tercile of sample distribution) and below normal (i.e. below lower tercile of sample distribution) electricity loads are considered in the forthcoming analysis. This notion of value is conceptually different from the notion of skill in the meteorological sense. In fact, the PEV cannot be assessed by analyzing meteorological variables alone, whereas it also depends on the users economic parameters. To evaluate the usefulness for end-users, in this study we assess the PEV of the electricity load forecast obtained using the grand ENSEMBLES-APCC/CliPAS MME and discuss the maximum PEV attainable as compared to the results using either ENSEMBLES or APCC/CliPAS MMEs. The generalized regression method described in De Felice et al. (2015) is applied to predict energy loads by exploiting all the information in the predictor (i.e. 2m temperature forecasts from the Grand ENSEMBLES-APCC/CliPAS MME). The evaluation of the PEV is performed using retrospective seasonal predictions of 2m temperature to forecast averaged June-July electricity demand during the period 1990-2007. All the results shown below have been obtained through a leave-one-out cross-validation procedure, where each year to be predicted is left out of the model-training sample.

Figure 13a shows the maximum PEV, as a function of the cost-loss ratio (C/L), obtained by applying the forecasts of electricity loads for the prediction of the binary events of getting above normal (upper panel) and below normal (lower panel) electricity-load outcomes. The Grand ENSEMBLES-APCC/CliPAS MME improves considerably compared to APCC/CliPAS and ENSEMBLES MMEs in particular for the upper tercile forecasts. Interestingly, the better combinations of APCC/CliPAS models have larger PEV compared to the better model combinations of ENSEMBLES, even if the performance of the best single-model is quite similar between the two MMEs for each C/L ratio (not shown). Therefore, the larger PEVs in APCC/CliPAS appear most likely related to the larger number of SPSs (11 vs. 5), which may increase the chance to introduce marginal contributions to performance (see section 2.1) for most values of the C/L ratio.

Each user is expected to have different C/L ratios and, in general, users with lower C/L will benefit more from the forecasts information by acting at lower probability thresholds. Although little is known about real-world costs and losses, economic considerations suggest that lower values of C/L are more likely than higher values (Roebber and Bosart, 1996) as indicated by the studies that have applied the simple cost-loss model to financial decisions (Thornes and Stephenson, 2001). We therefore consider the PEV averaged in the 0-0.3 C/L range. Remarkably, the PEV averaged over the 0-0.3 C/L range, reported in Figure 13b are largely positive. The Grand MME improves significantly (5% significance level) compared to the contributing ENSEMBLES and APCC/CliPAS MMEs. For the upper (lower) tercile the Grand-MME improves compared to APCC/CliPAS by 16% (15%) and compared to ENSEMBLES by 28% (25%).

4 Conclusions

The averaged performance of APCC/CliPAS and ENSEMBLES are comparable at the global-scale, but regionally the two systems appear to perform differently. The regional differences between APCC/CliPAS and ENSEMBLES indicate a high degree of independence for the two MMEs, which is a prerequisite to obtain skill amplification and error cancellations if combined.

Significant improvement of the skill (deterministic and probabilistic) is obtained over Tropics, northern middle latitudes (NML; 25N-75N) and southern middle latitudes (SML; 75S-25S) by collecting and exploiting all models into the Grand MME. In general, only a limited number of SPSs is required to obtain the maximum performance. The number and selection of models that perform better is usually different depending on the region/phenomenon under consideration so that all models are useful in some cases. Each model has its own distinction and provides added value for some region, season or variable. The analysis of all the possible multi-model combinations obtained by putting together the 5 models from ENSEMBLES and the 11 models from APCC/CliPAS confirms that realistic mean state of SST over the Tropics is a key aspect for the models in order to be able to simulate/predict interannual climate variability. It is shown a relationship between ensemble-mean performance (Pattern Correlation Coefficient; PCC) and tropical SST bias over both Tropics itself as well as over NML and SML regions.

The maximum probabilistic performance, obtained by identifying the better combination of models from the Grand MME, improves considerably over Tropics, NML and SML compared with the maximum performance attainable by using either the models from APCC/CliPAS only or from ENSEMBLES only. The averaged boreal-summer improvement over Tropics is 8.4%

(9.8%) for 2m temperature (precipitation), while it always exceed 10% over middle latitudes with 10.6% (10.8%) for NML and 10.5% (10.2%) for SML. In agreement with previous works suggesting that MMEs can be more effective in enhancing performance when combining SPSs developed by relatively independent communities, our results show that the incremental probability performance contribution tends to be higher when adding one model from ENSEMBLES to the APCC/CliPAS MME combinations and vice versa. Indeed it is shown that by adding ENSEMBLES (APCC/CliPAS) SPSs to the combinations of APCC-only (ENSEMBLES-only) models improves significantly (10% significance level) more than adding APCC/CliPAS (ENSEMBLES) in 86% (95%), 89% (96%), 63% (92%) of the times over NML, Tropics and SML, respectively. The increasing diversity of the Grand MME is therefore supposed to drive the improvements of the performance compared to the contributing MMEs. It is pointed out that APCC/CliPAS and ENSEMBLES MMEs may add diversity to each other both in terms of model formulation/parameterization and in terms of initialisation. Remarkably, it is noted that European models tend to use ECMWF reanalyses to start the hindcasts, whereas Asian-Pacific models use NCEP reanalyses in most cases. It is out of the scope of this paper to evaluate the diversity of the SPSs in APCC/CliPAS with respect to ENSEMBLES in their full complexity. Future works will be needed to identify what sort of diversity can contribute the most when trying to maximize MME performance.

The Euro-Mediterranean is one of the regions where the Grand MME improves significantly the probabilistic performance of temperature forecasts, so it is particularly meaningful to consider the application to the prediction of energy demand over Italy (as provided by Italian TSO, TERNA SpA) that is particularly sensitive to temperatures in summer. Indeed, the better combinations from the Grand MME produce a significant enhancement in the potential Economic Value (PEV) of the MME forecasts. The prediction of above (below) upper (lower) tercile energy demand for June-July improves by 15% and 36% (10% and 25%) with respect to the maximum PEV attainable from either ENSEMBLES or APCC/CliPAS, respectively. This demonstrates the potential of the Grand MME to contribute in obtaining useful predictions of electricity load at the seasonal time scale.

The results of the present study indicate that exploiting together the MMEs independently developed by the different communities is the way forward to optimize performance of seasonal climate predictions and to maximize the benefit for the end-users. It is recommended that the real-time multi-model ensembles that has been established as part of the operational seasonal forecast suites by the European (EUROSIP), the Asian-Pacific (APCC) and the North American (NMME) communities will be exploited together in the future to go beyond current limitations and pursue

480 increasingly useful climate predictions at the seasonal time-scale.

481 *Acknowledgements.* This work was supported by the European Union Seventh Framework Programme
482 (FP7/2007-13) under Grant 308378 (SPECS Project; http://specs-fp7.eu/) and under grant agreement N.
483 303208 (CLIMITS Project). Further support was provided to this work by the European Union's Hori-
484 zon 2020 research and innovation programme under grant agreement N. 641816 (CRESCENDO project;
485 http://crescendoproject.eu/) and under grant agreement N. 704585 (PROCEED project).

References

Adler, R. F., and Coauthors, 2003: The version-2 global precipitation climatology project (gpcp) monthly precipitation analysis (1979-present). *Journal of Hydrometeorology*, **4 (6)**, 1147–1167.

Ahn, J.-B., and H.-J. Kim, 2014: Improvement of 1-month lead predictability of the wintertime ao using a realistically varying solar constant for a cgcm. *Meteorological Applications*, **21 (2)**, 415–418, doi: 10.1002/met.1372, URL http://dx.doi.org/10.1002/met.1372.

Alessandri, A., A. Borrelli, S. Masina, A. Cherchi, S. Gualdi, A. Navarra, P. Di Pietro, and A. F. Carril, 2010: The ingv-cmcc seasonal prediction system: Improved ocean initial conditions. *Monthly Weather Review*, **138 (7)**, 2930–2952.

Alessandri, A., A. Borrelli, A. Navarra, A. Arribas, M. Déqué, P. Rogel, and A. Weisheimer, 2011a: Evaluation of probabilistic quality and value of the ENSEMBLES multi-model seasonal forecasts: comparison with DEMETER. *Mon. Wea. Rev.*, **139(2)**, 581–607, doi:10.1175/2010MWR3417.1.

Alessandri, A., A. Borrelli, A. Navarra, A. Arribas, M. Déqué, P. Rogel, and A. Weisheimer, 2011b: Evaluation of Probabilistic Quality and Value of the ENSEMBLES Multimodel Seasonal Forecasts: Comparison with DEMETER. *Monthly Weather Review*, **139 (2)**, 581–607, doi:10.1175/2010MWR3417.1.

Balmaseda, M., A. Vidard, and D. Anderson, 2008: The ecmwf ora-s3 ocean analysis system. *Mon. Wea. Rev*, **136 (8)**, 3018–3034.

Berrisford, P., D. Dee, K. Fielding, M. Fuentes, P. Kallberg, S. Kobayashi, and S. Uppala, 2009: The ERA-Interim archive. *ERA report series*, **1 (1)**.

Collins, W., and Coauthors, 2008: Evaluation of the hadgem2 model. *Hadley Cent. Tech. Note*, **74**.

Daget, N., A. Weaver, and M. Balmaseda, 2009: Ensemble estimation of background-error variances in a three-dimensional variational data assimilation system for the global ocean. *Quarterly Journal of the Royal Meteorological Society*, **135 (641)**, 1071–1094.

De Felice, M., A. Alessandri, and F. Catalano, 2015: Seasonal climate forecasts for medium-term electricity demand forecasting. *Applied Energy*, **137**, 435–444, doi:10.1016/j.apenergy.2014.10.030, URL http://linkinghub.elsevier.com/retrieve/pii/S030626191401071X.

Dubus, L., 2010: Practices, needs and impediments in the use of weather/climate information in the electricity sector. *Management of Weather and Climate Risk in the Energy Industry*, 175–188.

Fu, X., and B. Wang, 2004: The boreal-summer intraseasonal oscillations simulated in a hybrid coupled atmosphere-ocean model*. *Monthly Weather Review*, **132 (11)**, 2628–2649.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005a: The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. *Tellus*, **57A**, 219–233.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005b: The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. *Tellus, Series A: Dynamic Meteorology and Oceanography*, **57 (3)**, 219–233, doi:10.1111/j.1600-0870.2005.00103.x.

Jeong, H.-I., and Coauthors, 2012: Assessment of the apcc coupled mme suite in predicting the distinctive

climate impacts of two flavors of enso during boreal winter. *Climate Dynamics*, **39 (1)**, 475–493, doi: 10.1007/s00382-012-1359-3, URL http://dx.doi.org/10.1007/s00382-012-1359-3.

Jeong, J.-H., C.-H. Ho, D. Chen, and T.-W. Park, 2008: Experimental 6-month hindcast and forecast simulation using ccsm3. *APCC 2008 Technical Report*, **1**.

Jin, E. K., and Coauthors, 2008: Current status of enso prediction skill in coupled ocean-atmosphere models. *Clim. Dyn.*, **31**, 647, doi:10.1007/s00382-008-0397-3.

Keenlyside, N., M. Latif, M. Botzet, J. Jungclaus, and U. Schulzweida, 2005: A coupled method for initializing el nino southern oscillation forecasts using sea surface temperature. *Tellus A*, **57 (3)**, 340–356.

Kirtman, B. P., and Coauthors, 2014: The north american multimodel ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Am. Meteor. Soc.*, **95**, 585–601, doi:10.1175/BAMS-D-12-00050.1.

Kug, J.-S., I.-S. Kang, and D.-H. Choi, 2008: Seasonal climate predictability with tier-one and tier-two prediction systems. *Climate Dynamics*, **31 (4)**, 403–416.

Lee, D. Y., J.-B. Ahn, and J.-H. Yoo, 2015: Enhancement of seasonal prediction of east asian summer rainfall related to western tropical pacific convection. *Climate Dynamics*, **45 (3)**, 1025–1042, doi:10.1007/s00382-014-2343-x, URL http://dx.doi.org/10.1007/s00382-014-2343-x.

Lee, J.-Y., B. Wang, Q. Ding, K.-J. Ha, J.-B. Ahn, A. Kumar, B. Stern, and O. Alves, 2011: How predictable is the northern hemisphere summer upper-tropospheric circulation? *Climate Dynamics*, **37 (5)**, 1189–1203, doi:10.1007/s00382-010-0909-9, URL http://dx.doi.org/10.1007/s00382-010-0909-9.

Lee, J.-Y., and Coauthors, 2010: How are seasonal prediction skills related to models' performance on mean state and annual cycle? *Clim. Dyn.*, **35**, 267–283, doi:10.1007/s00382-010-0857-4.

Luo, J.-J., S. Masson, S. Behera, S. Shingu, and T. Yamagata, 2005: Seasonal climate predictability in a coupled oagcm using a different approach for ensemble forecasts. *Journal of Climate*, **18 (21)**, 4474–4497.

Mélia, D. S., 2002: A global coupled sea ice–ocean model. *Ocean Modelling*, **4 (2)**, 137–172.

Merryfield, W. J., and Coauthors, 2013: The canadian seasonal to interannual prediction system. part i: Models and initialization. *Monthly Weather Review*, **141 (8)**, 2910–2945.

Min, Y.-M., V. N. Kryjov, and S. M. Oh, 2014: Assessment of apcc multimodel ensemble prediction in seasonal climate forecasting: Retrospective (19832003) and real-time forecasts (2008-2013). *Journal of Geophysical Research: Atmospheres*, **119 (21)**, 12,132–12,150, doi:10.1002/2014JD022230, URL http://dx.doi.org/10.1002/2014JD022230, 2014JD022230.

Min, Y.-M., V. N. Kryjov, and S. M. Oh, 2017: Skill of real-time operational forecasts with the apcc multi-model ensemble prediction system during the period 2008-2015. *Climate Dynamics*, doi:10.1007/s00382-017-3576-2.

Palmer, T., and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal to interannual prediction (DEMETER). *Bull. Am. Meteor. Soc.*, **85**, 853–872.

Richardson, D., 2003: Economic value and skill. forecast verification: A practitioner's guide in atmospheric

559 science. i. joliffe and d. stephenson. wiley edition.
560 Richardson, D., 2006: Predictability and economic value. *Predictability of Weather and Climate*, T. Palmer,
561 and R. Hagedorn, Eds., Cambridge University Press, 628–644.
562 Roebber, P. J., and L. F. Bosart, 1996: The complex relationship between forecast skill and forecast value:
563 A real-world analysis. *Weather and Forecasting*, **11 (4)**, 544–559.
564 Rothstein, B., and G. Halbig, 2010: Weather Sensitivity of Electricity Supply and Data Services of the
565 German Met Office. *Management of Weather and Climate Risk in the Energy Industry*, 253–266, URL
566 http://www.springerlink.com/index/W547737515238333.pdf.
567 Saha, S., and Coauthors, 2006: The ncep climate forecast system. *Journal of Climate*, **19 (15)**, 3483–3517.
568 Thornes, J. E., and D. B. Stephenson, 2001: How to judge the quality and value of weather forecast products.
569 *Meteorological Applications*, **8 (03)**, 307–314.
570 Vintzileos, A., M. M. Rienecker, M. J. Suarez, S. K. Miller, P. J. Pegion, and J. T. Bacmeister, 2004: Sim-
571 ulation of the el niño–southern oscillation phenomenon with nasa's seasonal-to-interannual prediction
572 project coupled general circulation model. *Exchanges*, 25.
573 Vitart, F., and Coauthors, 2007: Dynamically-based seasonal forecasts of atlantic tropical storm activity
574 issued in june by eurosip. *Geophys. Res. Lett.*, **34**, L16 815, doi:doi:10.1029/2007GL030740.
575 Wang, B., and Coauthors, 2009: Advance and prospectus of seasonal prediction: assessment of the
576 apcc/clipas 14-model ensemble retrospective seasonal prediction (1980–2004). *Climate Dynamics*,
577 **33 (1)**, 93–117.
578 Weigel, A., M. Liniger, and C. Appenzeller, 2008: Can multi-model combination really enhance the pre-
579 diction skill of probabilistic ensemble forecasts? *QJRMS*, **134**, 241–260.
580 Weisheimer, A., and Coauthors, 2009: Ensembles: A new multi-model ensemble for seasonal-to-annual
581 predictions—skill and progress beyond demeter in forecasting tropical pacific ssts. *Geophysical research*
582 *letters*, **36 (21)**.
583 Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*, Vol. 100. Academic Press.
584 Yoo, J. H., and I.-S. Kang, 2005: Theoretical examination of a multi model composite. *Geophy. Res. Lett.*,
585 **32**, L18 707, doi:10.1029/2005GL023513.
586 Zhang, S., M. Harrison, A. Rosati, and A. Wittenberg, 2007: System design and evaluation of coupled
587 ensemble data assimilation for global oceanic climate studies. *Monthly Weather Review*, **135 (10)**, 3541–
588 3564.
589 Zhong, A., H. H. Hendon, and O. Alves, 2005: Indian ocean variability and its association with enso in a
590 global coupled model. *Journal of Climate*, **18 (17)**, 3634–3649.

Tables

APCC/CliPAS	ENSEMBLES
APCC Asia-Pacific Economic Cooperation Climate Center, S. Korea. (Jeong et al., 2008, 2012)	ECMWF, European Centre for Medium-Range Weather Forecasts, United Kingdom (Balmaseda et al., 2008)
NCEP, National Center for Environmental Prediction, USA (Saha et al., 2006)	UKMO, UK-Met Office Met Office, United Kingdom (Collins et al., 2008)
BMRC, Bureau of Meteorology Research Center, Australia (Zhong et al., 2005)	MF, Meteo France. France (Mélia, 2002; Daget et al., 2009)
PNU, Pusan National University, S. Korea. (Ahn and Kim, 2014)	INGV-CMCC, Centro Euro-Mediterraneo per i Cambiamenti Climatici, Italy (Alessandri et al., 2010)
MSC, Meteorological Service of Canada, Canada (CANCM3, CANCM4) (Merryfield et al., 2013)	IFM-GEOMAR, Leibnitz Institute of Marine Sciences at Kiel University, Germany (Keenlyside et al., 2005)
NASA, National Aeronautics and Space Administration, USA (Vintzileos et al., 2004)	
SNU, Seoul National University, S. Korea (Kug et al., 2008)	
UH, University of Hawaii, USA (Fu and Wang, 2004)	
GFDL, The Geophysical Fluid Dynamics Laboratory , USA (Zhang et al., 2007)	
FRCGC, Frontier Research Center for Global Change, Japan (Luo et al., 2005)	

Table 1: Contributing institutions to the grand ENSEMBLES-APCC/CliPAS MME

Institute	AGCM	OGCM	Members (JJA/DJF)	Initialization	
				Atmos. & Land	Ocean
APCC/CLiPAS					
APCC	CAM3 (T85/L26)	POP1.3 (gx1v3/L40)	10	Coupled atmos-ocean climate simulations with SST restored to observations	Derived from NCEP ocean reanalysis (GODAS)
NCEP	GFS (T126/L64)	MOM3 (0.33°× 1°/L40)	20	NCEP CFS Data Assimilation	NCEP ocean reanalysis (GODAS)
BMRC	BAM3.0d (T47/L17)	ACOM2 (0.5°-1.5°× 2°/L25)	30	AMIP-type simulation with forced SSTs	Off-line ocean analysis. NCEP surface fluxes except for wind-stress from FSU
PNU	CCM3 (T42/L18)	MOM3 (0.7°-2.8°/L29)	10/5	AMIP-type simulation with forced SSTs	Off-line ocean analysis. NCEP surface fluxes
MSC_CANCM3	AGCM3 (T63/L31)	OGCM4 (0.94°× 1.41°/L40)	10	Canadian Meteorological Centre atmospheric assimilation	SST and sea-ice analysis (surface assimilation). Below surface 3D analysis based on NCEP GODAS
MSC_CANCM4	AGCM4 (T63/L35)	OGCM4 (0.94°× 1.41°/L40)	10	Canadian Meteorological Centre atmospheric assimilation	SST and sea-ice analysis (surface assimilation). Below surface 3D analysis based on NCEP GODAS
NASA	GEOS5 (288x181/L72)	MOM4 (720 × 410/L40)	9/11	MERRA atmospheric reanalysis	Ocean analysis from GMAO ODAS
SNU	SNU (T42/L21)	MOM2.2 (0.33°× 1°/L32)	6	Atmos. and land IC obtained from NCEP reanalysis	Off-line ocean analysis
UH	ECHAM4 (T31/L19)	UH Ocean (1°× 2°/L2)	10	coupled atmos-ocean climate simulations with SST and thermocline depth restored to observations	
GFDL	AM2.1 (2°× 2.5°/L24)	OM3.1/MOM4 (0.33°× 1°/L50)	10	coupled atmos-ocean data assimilation	
FRCGC	ECHAM4 (T106/L19)	OPA8.2 (2°× 2°/L31)	9	coupled atmos-ocean climate simulations with SST restored to observations	
ENSEMBLES					
ECMWF	IFS CY31R1 (T159/L62)	HOPE (0.3°-1.4°/L29)	9	ERA40/ECMWF oper. analysis	Ocean analysis forced by ERA-40/ECMWF oper. analysis surface fluxes
UKMO	HadGEM2-A (N96/L38)	HadGEM2-O (0.33°-1°/L20)	9	ERA40/ECMWF oper. analysis, anomaly assimilation for soil moisture	Ocean analysis forced by ERA-40/ECMWF oper. analysis surface fluxes
MF	ARPEGE4.6 (T63/L31)	OPA8.2/GELATO (2°/L31)	9	Atmos. nudging to ERA-40/ECMWF oper. analysis	Ocean analysis forced by ERA-40/ECMWF oper. analysis surface fluxes
INGV-CMCC	ECHAM5-SILVA (T63/L19)	OPA8.2 (2°/L31)	9	AMIP-type simulation with forced SSTs	Ocean analysis forced by ERA-40/ECMWF oper. analysis surface fluxes
IFM-GEOMAR	ECHAM5 (T63/L31)	MPI-OM1 (1.5°/L40)	9	coupled atmos-ocean climate simulations with SST restored to observations	

Table 2: Model configuration, resolution, ensemble members provided and initialization strategy of each institution (see reference papers in Table 1).

	season	GrandMME vs ENSEMBLES	GrandMME vs APCC/CliPAS
t2m	DJF	5% (5%)	5% (6%)
	JJA	5% (7%)	5% (8%)

Table 3: The ratio (%) of improvement in the globally-averaged correlation (sea and land grid points) for the Grand MME compared with (left) ENSEMBLES and (right) APCC/CliPAS. Brackets indicate results for land-only grid points.

variable	area	null hypothesis distribution		Actual improvement
		p05	p95	
		JJA		
t2m	NML	0.006	0.030	0.032 *
	TROPICS	-0.001	0.021	0.024 *
	SML	0.008	0.040	0.041 *
prec	NML	-0.000	0.020	0.023 *
	TROPICS	0.005	0.023	0.025 *
	SML	0.002	0.014	0.012

Table 4: Actual Grand MME enhancements in normalized marginal contributions, as from Figure 12, together with the 5th and 95th percentiles of the null hypothesis distribution of getting as high or higher improvements just as a consequence of the larger number of model combinations. For each domain, asterisk indicate that improvements are significant at the 5% level.

variable	area	JJA	DJF
t2m	NML	10.6	8.6
	TROPICS	8.4	7.0
	SML	10.5	10.3
prec	NML	10.8	11.5
	TROPICS	9.8	10.4
	SML	10.1	10.9

Table 5: The ratio (%) of improvement of the brier skill score averaged over NML (25N-75N), Tropics (25S-25N), and SML (75S-25S) for the grand MME compared with the maximum performance of the contributing ENSEMBLES or APCC/CliPAS MMEs (whichever is larger at each grid point) for (upper rows) 2m temperature and (lower rows) precipitation.

Figure Captions

Fig. 1. 1-month-lead boreal summer (JJA) 2m Temperature (a) correlation of ENSEMBLES with ERA-INTERIM, (b) ENSEMBLES minus APCC/CliPAS correlation difference and (c) frequency distribution of the ENSEMBLES minus APCC/CliPAS correlation differences. Dotted grid points in (a) and (b) did pass a significance test at 5% level.

Fig. 2. (a) grand ENSEMBLES-APCC/CliPAS minus ENSEMBLES MMEs difference of the correlation for JJA 2m-temperature seasonal forecasts at one month of lead time. (b) distribution of the point-by-point differences between the grand ENSEMBLES-APCC/CliPAS and ENSEMBLES both on land (light blue) and on all (sea and land) grid points (red). Vertical lines represent the average value of the distributions.

Fig. 3. (a) NML (25N-75N), (b) Tropics (25S-25N) and (C) SML (75S-25S) Pattern Correlation Coefficients (PCCs) for boreal summer (JJA) 2m temperature computed as a function of the number of models obtained with all the possible 65535 combinations of the models from ENSEMBLES (5 models) and APCC/CliPAS (11 models). Blue (red) colour represent the cases in which the combinations are obtained with only models from ENSEMBLES (APCC/CliPAS), while green colour is for the combinations of the models from both APCC/CliPAS and ENSEMBLES. The average of the combinations (filled circles) in each category is reported with the diamonds while the maximum PCC for each type are the dashed horizontal lines.

Fig. 4. Same as Figure 3 but for precipitation.

Fig. 5. PCCs for 2m temperature vs. Mean Tropical SST bias for (a) NML, (b) Tropics and (C) SML obtained with all the possible combinations of models coming from APCC/CliPAS and ENSEMBLES MMEs. The linear fit of all the MME combinations is reported when significance of the slope of linear relationship is verified at the 5% level using a Fisher parametric test.

Fig. 6. Same as Figure 5 but for precipitation

Fig. 7. (a) Maximum Brier Skill Score (BSS) that is attainable for 1-month lead seasonal forecasts started May 1st (JJA) using the grand ENSEMBLES-APCC/CliPAS MME. The BSS is obtained for each grid point by getting the maximum value of all the possible 65535 combinations of the 5 models from ENSEMBLES and the 11 models from APCC/CliPAS. (b) BSS % gain by using the grand ENSEMBLES-APCC/CliPAS MME by comparison with the maximum performance of ENSEMBLES and APCC/CliPAS.

Fig. 8. Same as Figure 7 but for precipitation

Fig. 9. Percent incremental contribution of each model to the BSS of the prediction of above normal 2m temperature in boreal summer (JJA) for (a) NML, (b) Tropics and (c) SML obtained by averaging the skill change of adding the given model to all 32767 possible combinations not already including it.

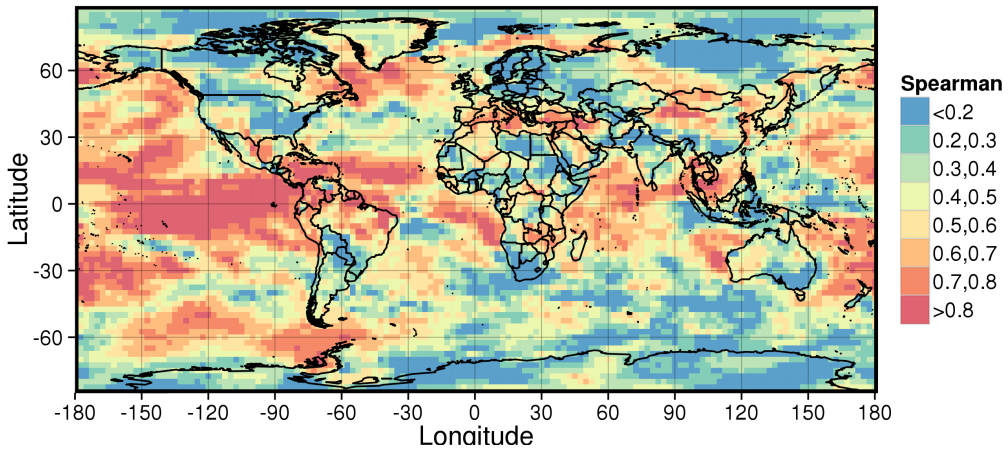
Fig. 10. Same as Figure 9 but for precipitation.

Fig. 11. Fraction of grid points considering the global domain where each model is needed in order to maximise BSS of the prediction of above normal (a) 2m temperature as reported in Figure 7 and (b) precipitation as reported in Figure 8. Colors indicate the number of models needed to maximize BSS for each relative fraction of grid points.

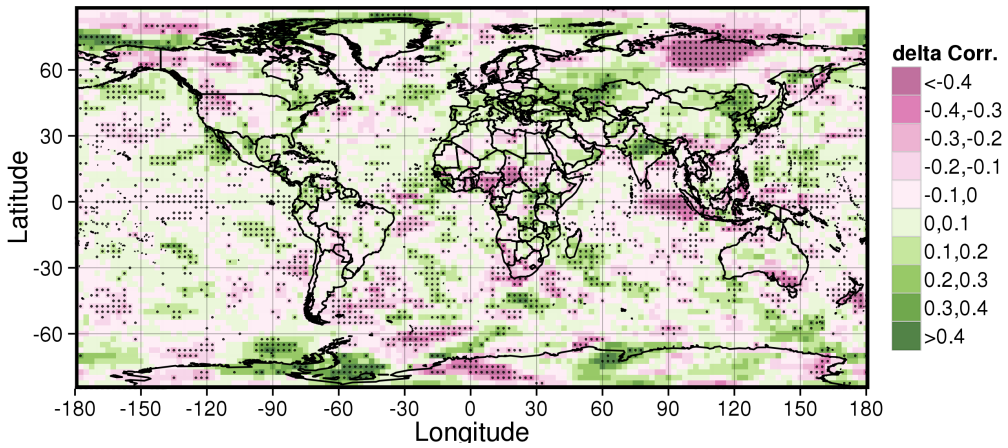
Fig. 12. Normalized marginal contribution of (red) APCC/CliPAS or (green) ENSEMBLES models to combinations of (left) APCC only, (right) ENSEMBLES only and (middle) mixed MMEs for (a) NML, (b) Tropics and (c) SML. The skill contributions are computed by averaging the skill change of adding APCC/CliPAS or ENSEMBLES models to all combinations (excluding combinations already including model to be added) in the APCC-only, ENSEMBLES-only and mixed categories.

Fig. 13. Potential economic value (PEV) of the grand ENSEMBLES-APCC/CliPAS (blue), ENSEMBLES (green) and APCC/CliPAS (red) forecasts for the prediction of June-July electricity load over Italy being (lower) below the lower tercile and (upper) above the upper tercile of the sample climatology (a) as a function of the C/L ratio and (b) averaged over the 0-0.3 C/L range.

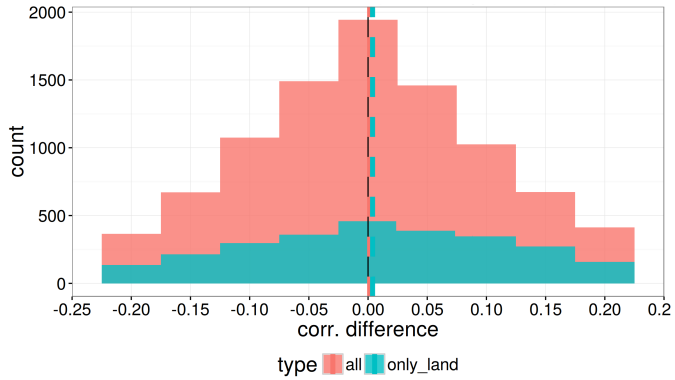
Figures



(a) Correlation of ENSEMBLES vs ERA-Interim T2m JJA

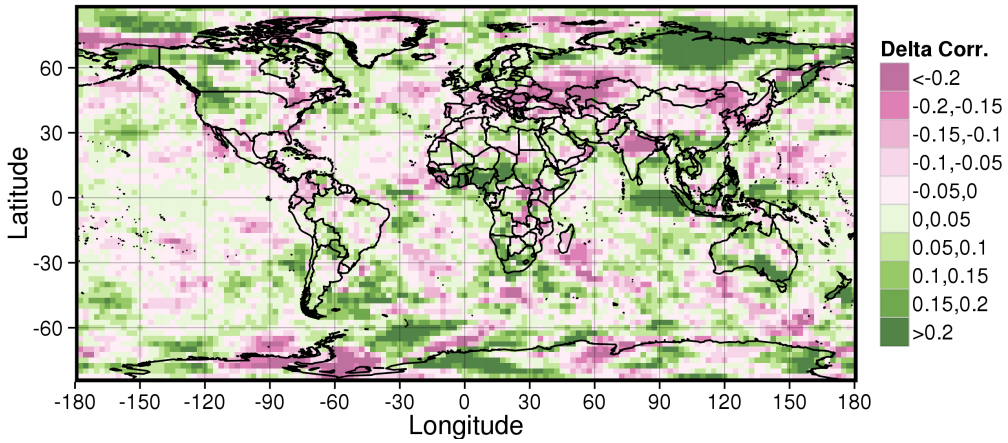


(b) ENSEMBLES minus APCC/CliPAS

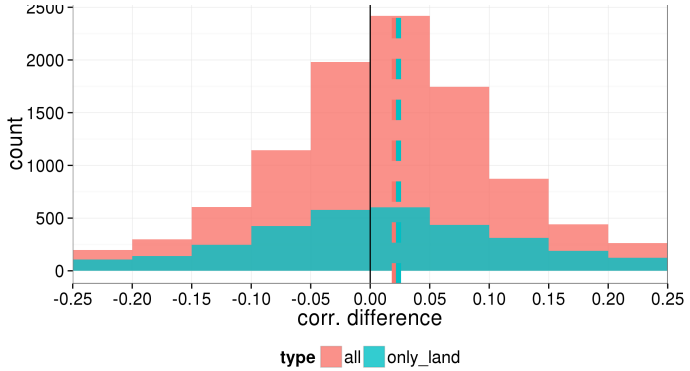


(c) Histogram of ENSEMBLES minus APCC/CliPAS differences

Fig. 1: 1-month-lead boreal summer (JJA) 2m Temperature (a) correlation of ENSEMBLES with ERA-INTERIM, (b) ENSEMBLES minus APCC/CliPAS correlation difference and (c) frequency distribution of the ENSEMBLES minus APCC/CliPAS correlation differences. Dotted grid points in (a) and (b) did pass a significance test at 5% level.

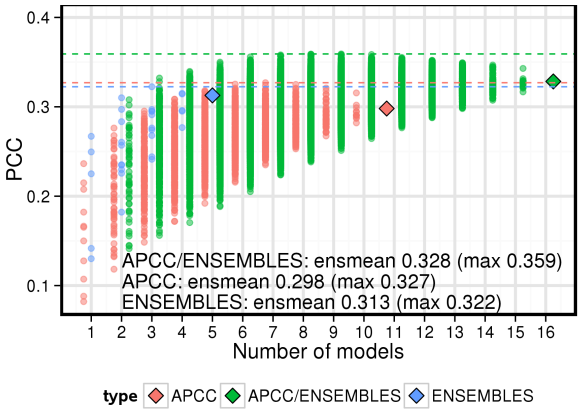


(a) Grand-MME minus ENSEMBLES T2m JJA

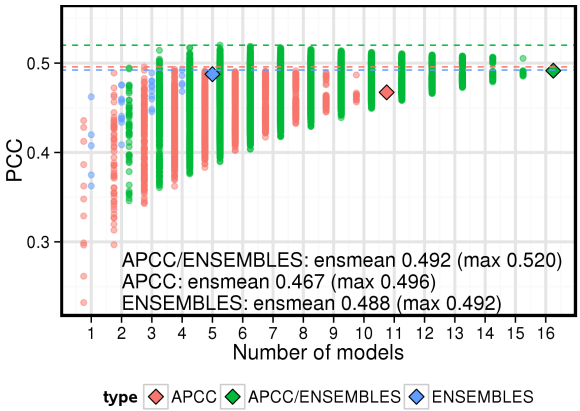


(b) Grand-MME minus ENSEMBLES Histogram T2m JJA

Fig. 2: (a) grand ENSEMBLES-APCC/CliPAS minus ENSEMBLES MMEs difference of the correlation for JJA 2m-temperature seasonal forecasts at one month of lead time. (b) distribution of the point-by-point differences between the grand ENSEMBLES-APCC/CliPAS and ENSEMBLES both on land (light blue) and on all (sea and land) grid points (red). Vertical lines represent the average value of the distributions.



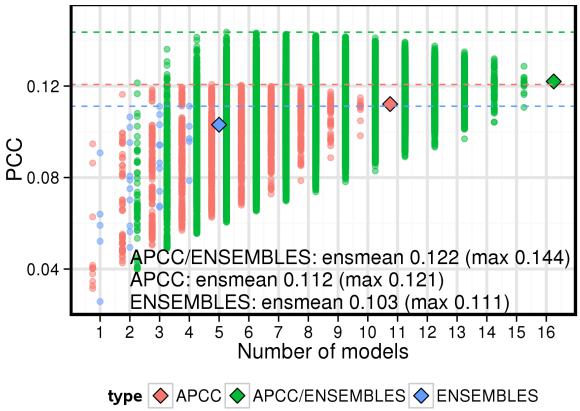
(a) NML T2m JJA



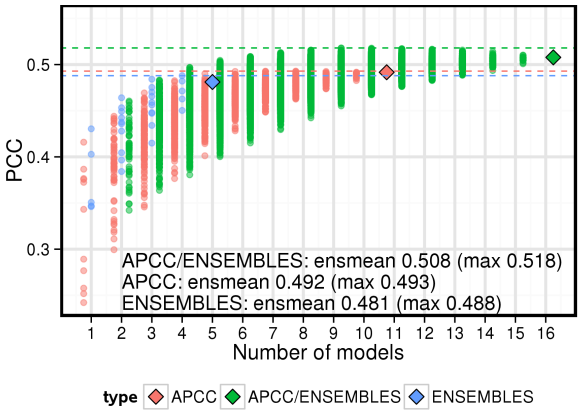
(b) TROPICS T2m JJA

(c) SML T2m JJA

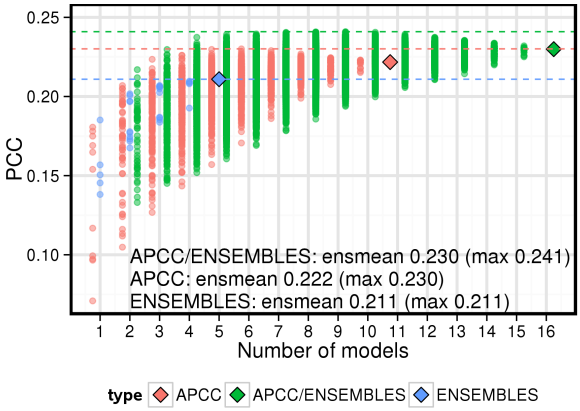
Fig. 3: (a) NML (25N-75N), (b) Tropics (25S-25N) and (C) SML (75S-25S) Pattern Correlation Coefficients (PCCs) for boreal summer (JJA) 2m temperature computed as a function of the number of models obtained with all the possible 65535 combinations of the models from ENSEMBLES (5 models) and APCC/CliPAS (11 models). Blue (red) colour represent the cases in which the combinations are obtained with only models from ENSEMBLES (APCC/CliPAS), while green colour is for the combinations of the models from both APCC/CliPAS and ENSEMBLES. The average of the combinations (filled circles) in each category is reported with the diamonds while the maximum PCC for each type are the dashed horizontal lines.



(a) NML Prec JJA

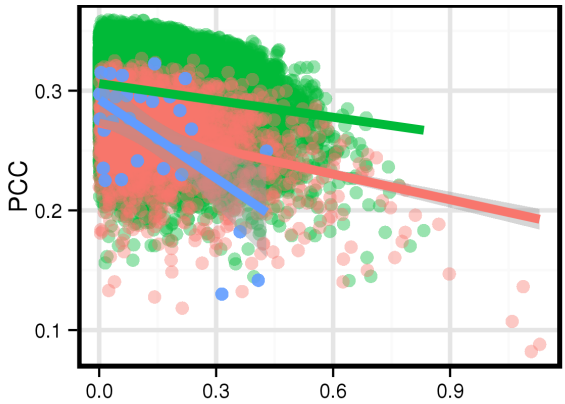


(b) TROPICS Prec JJA

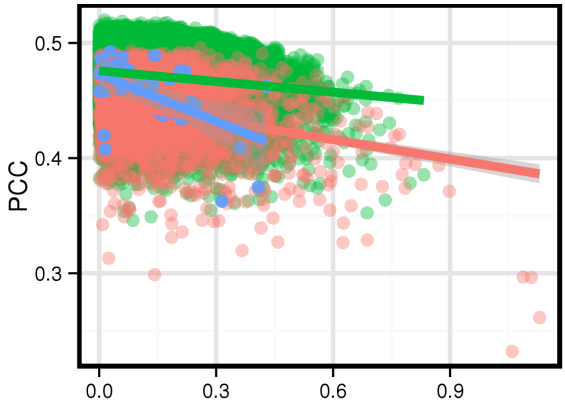


(c) SML Prec JJA

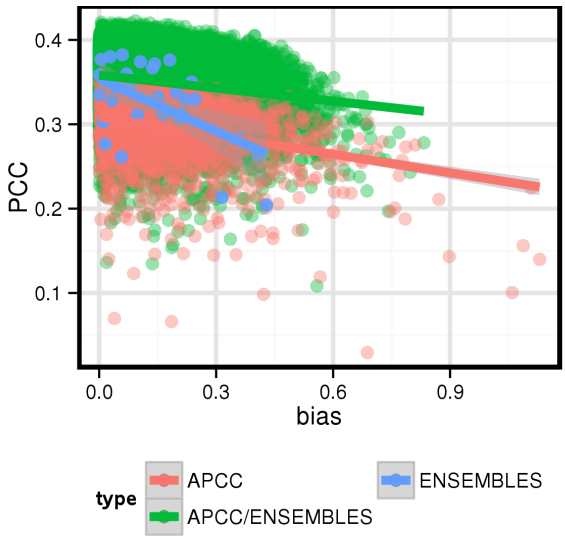
Fig. 4: Same as Figure 3 but for precipitation.



(a) NML PCC T2m vs. TROPICS SST bias JJA

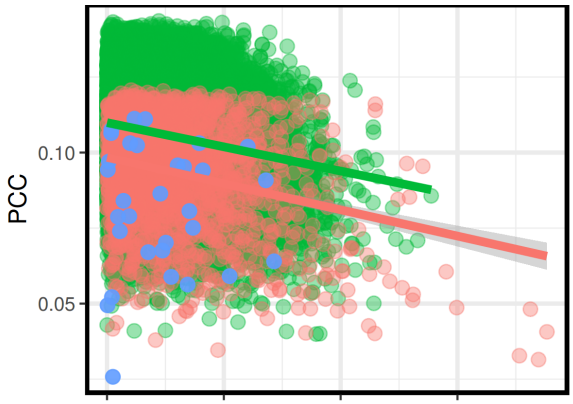


(b) TROPICS PCC T2m vs. TROPICS SST bias JJA

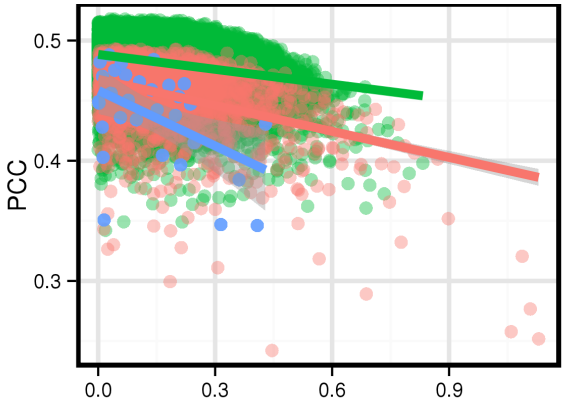


(c) SML PCC T2m vs. TROPICS SST bias JJA

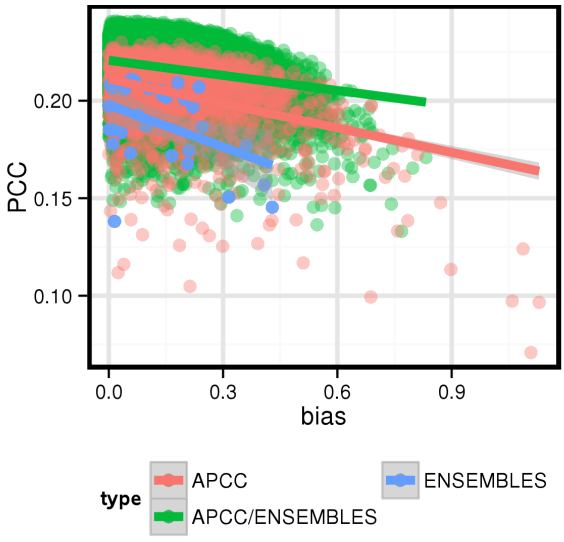
Fig. 5: PCCs for 2m temperature vs. Mean Tropical SST bias for (a) NML, (b) Tropics and (C) SML obtained with all the possible combinations of models coming from APCC/CliPAS and ENSEMBLES MMEs. The linear fit of all the MME combinations is reported when significance of the slope of linear relationship is verified at the 5% level using a Fisher parametric test.



(a) NML PCC Prec vs. TROPICS SST bias JJA

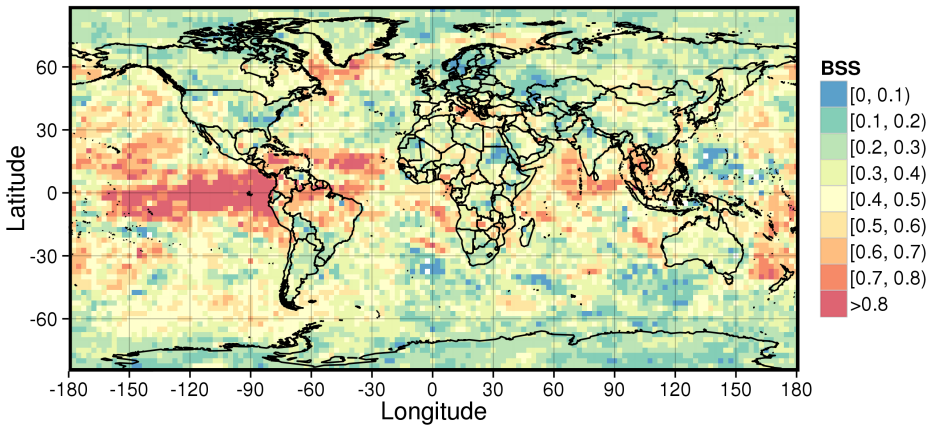


(b) TROPICS PCC Prec vs. TROPICS SST bias JJA

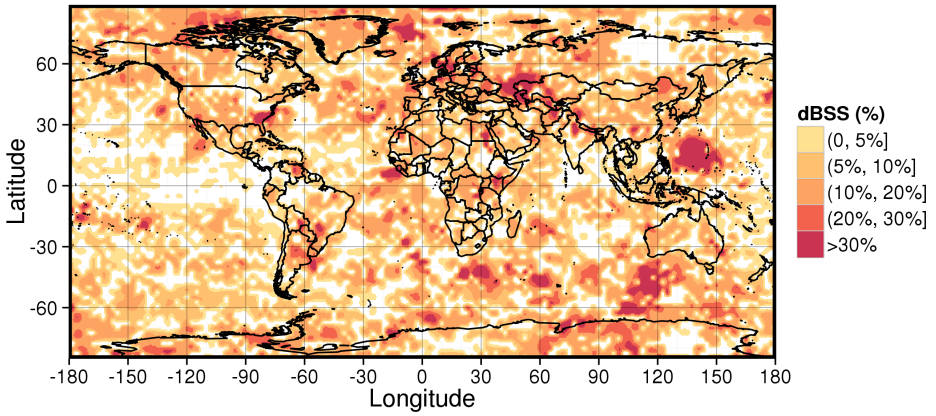


(c) SML PCC Prec vs. TROPICS SST bias JJA

Fig. 6: Same as Figure 5 but for precipitation

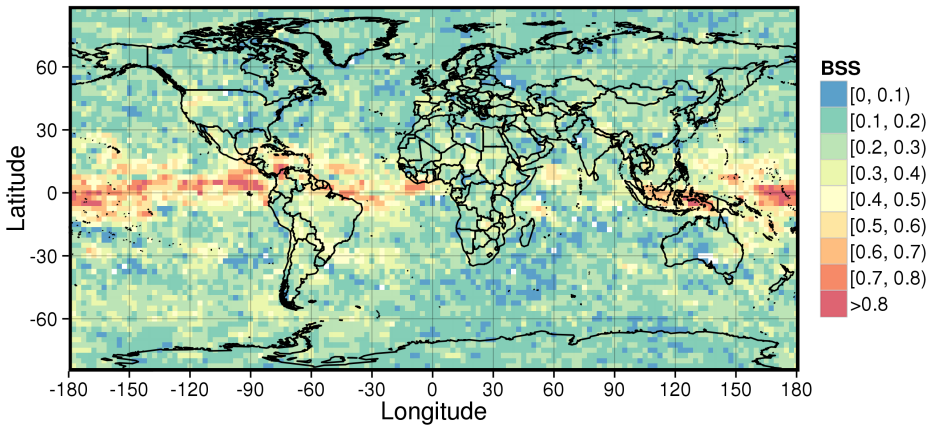


(a) Grand-MME T2m JJA BSS

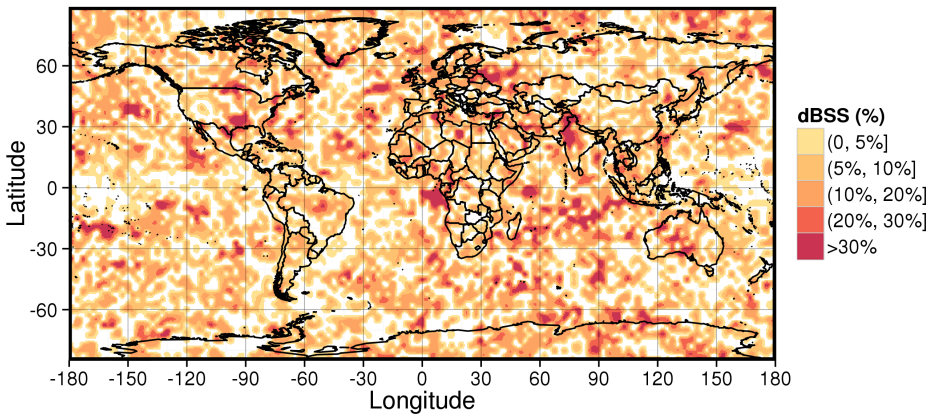


(b) Grand-MME minus Max(APCC/CliPAS, ENSEMBLES) – Delta %

Fig. 7: (a) Maximum Brier Skill Score (BSS) that is attainable for 1-month lead seasonal forecasts started May 1st (JJA) using the grand ENSEMBLES-APCC/CliPAS MME. The BSS is obtained for each grid point by getting the maximum value of all the possible 65535 combinations of the 5 models from ENSEMBLES and the 11 models from APCC/CliPAS. (b) BSS % gain by using the grand ENSEMBLES-APCC/CliPAS MME by comparison with the maximum performance of ENSEMBLES and APCC/CliPAS.

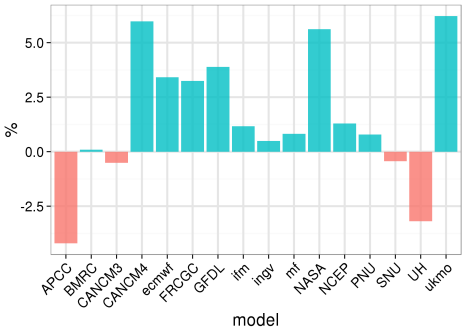


(a) Grand-MME Prec JJA BSS

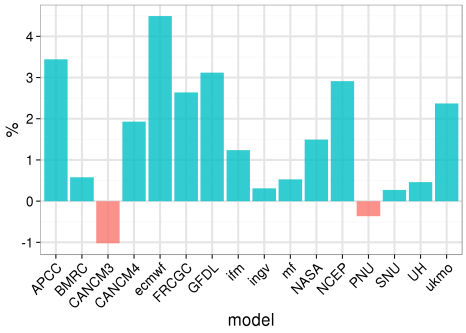


(b) Grand-MME minus Max(APCC/CliPAS, ENSEMBLES) – Delta %

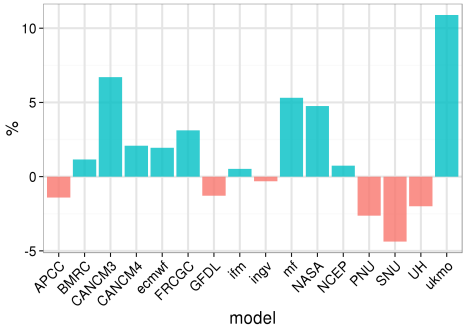
Fig. 8: Same as Figure 7 but for precipitation



(a) NML marginal % BSS T2m

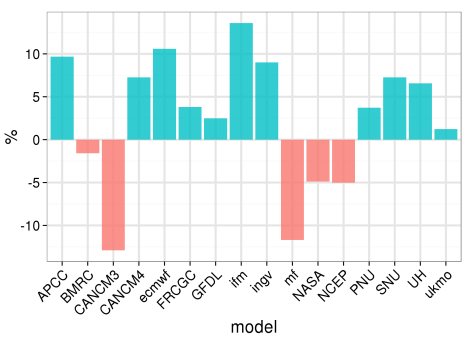


(b) TROPICS marginal % BSS T2m

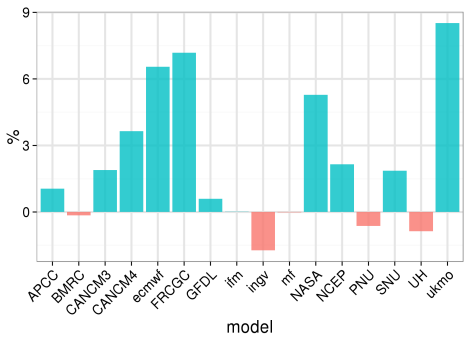


(c) SML marginal % BSS T2m

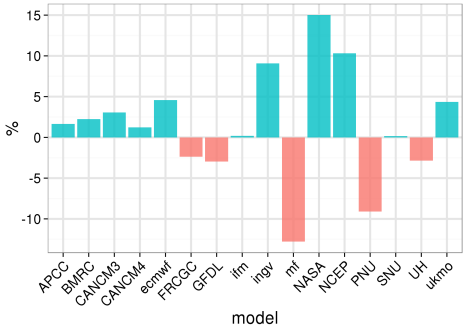
Fig. 9: Percent incremental contribution of each model to the BSS of the prediction of above normal 2m temperature in boreal summer (JJA) for (a) NML, (b) Tropics and (c) SML obtained by averaging the skill change of adding the given model to all 32767 possible combinations not already including it.



(a) NML marginal % BSS Prec

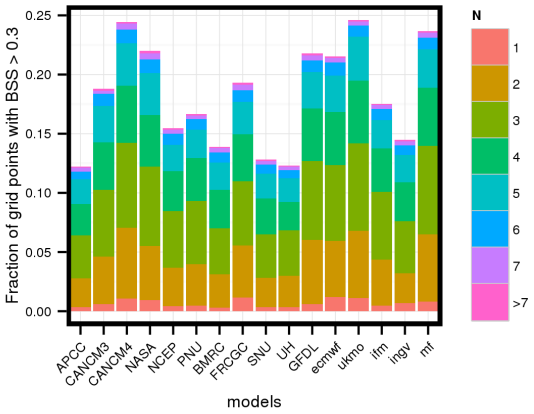


(b) TROPICS marginal % BSS Prec

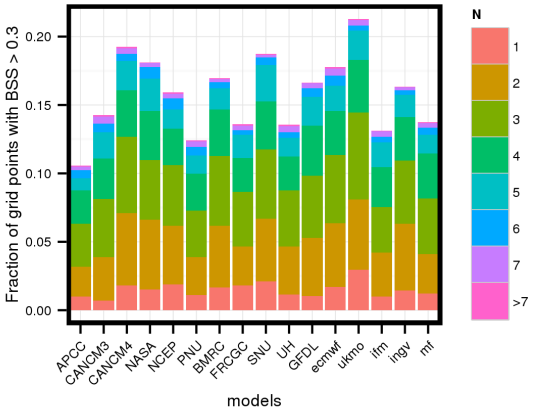


(c) SML marginal % BSS Prec

Fig. 10: Same as Figure 9 but for precipitation.



(a) SELECTION T2m JJA filter 0.3



(b) SELECTION Prec JJA filter 0.3

Fig. 11: Fraction of grid points considering the global domain where each model is needed in order to maximise BSS of the prediction of above normal (a) 2m temperature as reported in Figure 7 and (b) precipitation as reported in Figure 8. Colors indicate the number of models needed to maximize BSS for each relative fraction of grid points.

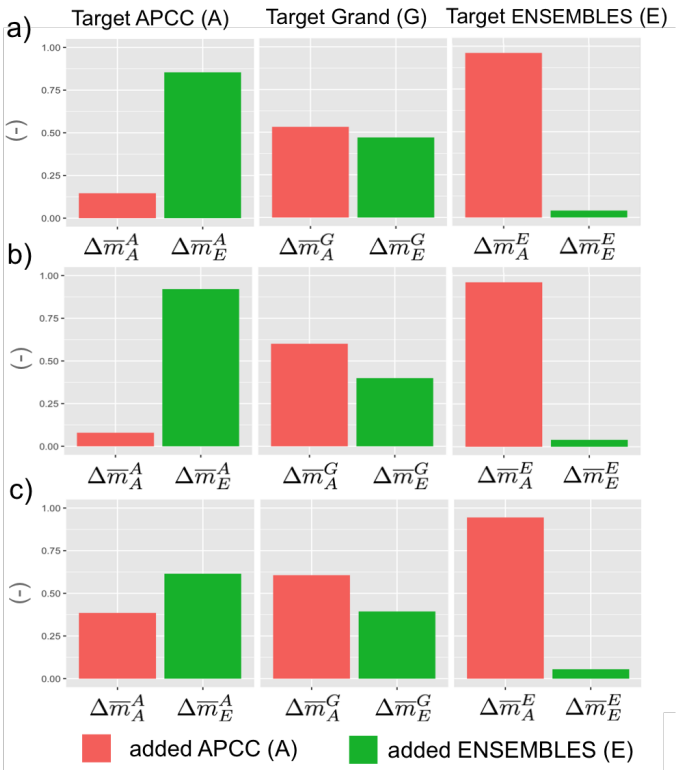


Fig. 12: Normalized marginal contribution of (red) APCC/CliPAS or (green) ENSEMBLES models to combinations of (left) APCC only, (right) ENSEMBLES only and (middle) mixed MMEs for (a) NML, (b) Tropics and (c) SML. The skill contributions are computed by averaging the skill change of adding APCC/CliPAS or ENSEMBLES models to all combinations (excluding combinations already including model to be added) in the APCC-only, ENSEMBLES-only and mixed categories.

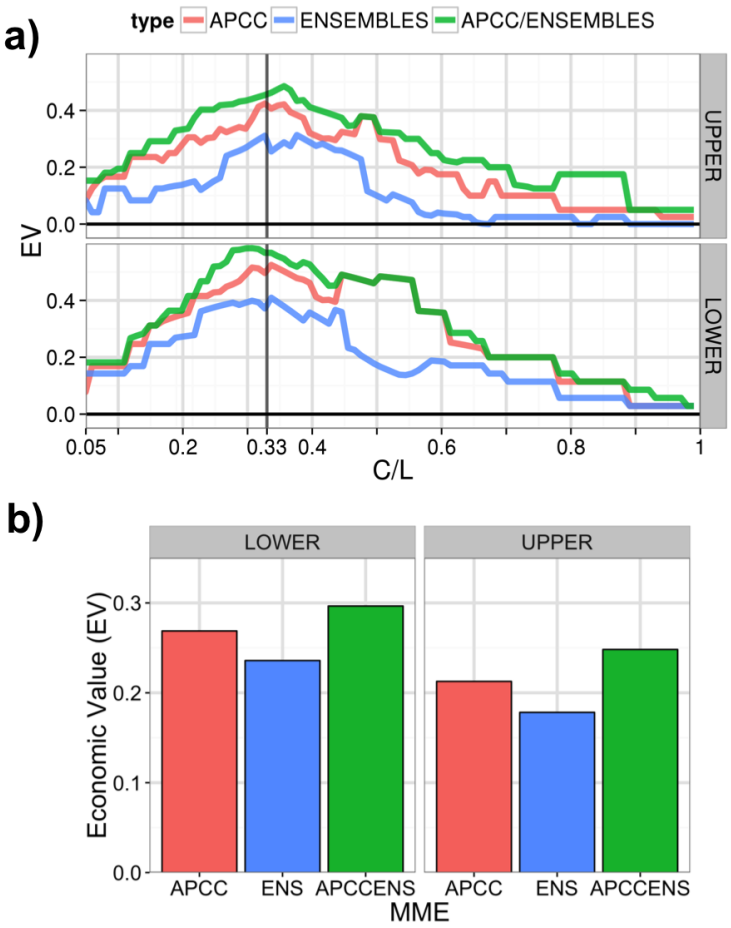


Fig. 13: Potential economic value (PEV) of the grand ENSEMBLES-APCC/CliPAS (blue), ENSEMBLES (green) and APCC/CliPAS (red) forecasts for the prediction of June-July electricity load over Italy being (lower) below the lower tercile and (upper) above the upper tercile of the sample climatology (a) as a function of the C/L ratio and (b) averaged over the 0-0.3 C/L range.