Grant agreement no. 243964

QWeCI

Quantifying Weather and Climate Impacts on Health in Developing Countries

D3.1.a

Assessment report on the skill of global seasonal predictions in Africa using a quintile interval-based validation

Start date of project: 1st February 2010 Duration: 42 months

Lead contractor : CSIC
Coordinator of deliverable : J.M.Gutiérrez Llorente
Evolution of deliverable

Due date : Month 9
Date of first draft : 01/01/2011
Start of review : 15/01/2011
Accepted by WP coordinator : 31/04/2011

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination Level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	X
CO	Confidential, only for members of the consortium (including the Commission Services)	

ASSESSMENT REPORT ON THE SKILL OF GLOBAL SEASONAL PREDICTIONS IN AFRICA USING A QUINTILE INTERVAL-BASED VALIDATION

R. MANZANAS[1], J.M. GUTIÉRREZ[1], M.D. FRÍAS[2], and A.S. COFIÑO[2]

Abstract. The skill of state-of-the-art seasonal accumulated precipitation and maximum temperature forecasts is assessed worldwide for the period 1961-2000 using the hindcasts provided by the project ENSEMBLES (Stream2 multi-model simulations). A quintile-based robust statistical validation is applied to one and four months lead-time predictions, obtaining the skill scores and their confidence intervals grid point by grid point. Results show that highest skill concentrates around the tropics for both variables; moreover, agreement among the models tend to be large in these regions. Overall, Autumn and Winter are the most skillful seasons for precipitation, whereas Winter and Spring are the most skillful for maximum temperature.

1. Introduction

Seasonal forecasting is a promising research field with enormous potential impact in different socio-economic sectors, including health [see *Kirtman and Pirani*, 2008, for a review]. Nowadays, several seasonal forecasting systems are run all around the world once or twice a month, providing weather anomalies a few months in advance. An example is the European EURO-SIP multi-model system, which is an operational system based on the research and development done in the EU-funded DEMETER [*Palmer et al.*, 2004] and ENSEMBLES [*Weisheimer et al.*, 2009] projects. In particular the ENSEMBLES dataset constitutes the longest to date state-of-the-art seasonal hindcast experiment.

Since seasonal predictability strongly vary from region to region and from season to season [see, e.g., *Halpert and Ropelewski*, 1992], a key task in this field is the appropriate assessment of the seasonal forecasting systems. However, only a few papers have been devoted to this problem, focusing always on particular models or regions and using different validation scores. For instance, a validation of the ENSEMBLES multi-model dataset for precipitation in three African regions has been recently published by [*Batte and Deque*, 2011], considering deterministic (e.g. ACC) and probabilistic scores (e.g. RPSS).

In the present deliverable, we assess the skill of state-of-the-art seasonal forecasts for precipitation and maximum temperature worldwide, using the ENSEMBLES multi-model dataset and applying a recently introduced quintile-based validation method which allows obtaining the skill scores and their statistical significance [*Frías et al.*, 2010; *Díez et al.*, 2011].

The deliverable is organized as follows: The data used is described in Sec. 2. The multi-model construction is justified in Sec. 4. The validation methodology is explained in Sec. 3. Finally, results are presented for precipitation (Sec. 5) and maximum temperature (Sec. 6). Main conclusions are summarized in Sec. 7.

[1]Instituto de Física de Cantabria (IFCA), CSIC-Universidad de Cantabria, Santander, Spain.

[2]Departamento de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, Santander, Spain.

0148-0227/11/$9.00

2. Data

In this section we describe the observed and predicted datasets used in this study, with a common period 1961-2000. In order to analyze different lead times (one- and four-months), we considered the Boreal seasons: Winter (DJF), Spring (MAM), Summer (JJA) and Autumn (SON); as we show below, for other seasons (e.g. West African Monsson, FMA) there is only a lead time to perform the analysis, thus limiting the analysis. In the case of precipitation, accumulated precipitation is used for the observed (predicted) values corresponding to each season; for maximum temperature, seasonal means of daily maximum temperatures are considered.

2.1. Observations

For monthly accumulated precipitation we use observations from VASClimO [see *Schneider et al.*, 2008, and http://gpcc.dwd.de], a worldwide quality controlled gridded dataset covering the period 1951-2000, at a 2.5° resolution.

For monthly maximum temperature we use observations from the CRU TS 2.1 dataset [see *Mitchell and Jones*, 2005, and http://www.cru.uea.ac.uk/cru/data/hrg/cru_ts_2.10/] from the Climate Research Unit. This dataset covers the period 1901-2002 and was bi-linearly interpolated to the same 2.5° resolution grid of the VASClimO dataset.

2.2. Models

We consider the seasonal predictions from the multi-model Stream2 experiment of the ENSEMBLES project (http://www.ecmwf.int/research/EU_projects/ENSEMBLES), comprising five state-of-the-art coupled atmosphere-ocean models from the following centers: The UK Met Office (UKMO), Météo France (MF), the European Centre for Medium-Range Weather Forecasts (ECMWF), the Leibniz Institute of Marine Sciences (IFM-GEOMAR) and the Euro-Mediterranean Centre for Climate Change (CMCC-INGV). Each of these models is formed by an ensemble of nine members. Seven months-long hindcasts were issued four times a year within the period 1960-2005, starting the first of February, May, August and November [see *Weisheimer et al.*, 2009, for more details about the experiment]. This allow us analysing one- and four-months lead-time predictions (i.e., initializations of August and May provide one- and four-month lead time forecasts for SON season, respectively). In order to have a common grid for both observations and predictions, models are bi-linearly interpolated to the 2.5° grid of the observations.

3. Methodology

The validation methodology used in this deliverable is the described in *Frías et al.* [2010], which is applied here worldwide grid point by grid point. For a particular season, observations and predictions are divided into five categories, according to their respective quintiles within the period 1961-2000. Then, a probabilistic forecast is computed year by year by considering the frequencies $p_i = n_i/n$ of the different quintiles $i = 1, 2, 3, 4, 5$, where n_i is the number of members falling within the the i-*th* quintile, out of a total of n members. For instance, $n = 9$ for each single model and $n = 45$ for the multi model combined assuming equal weights for all models and members. Working with order statistics (quintiles), instead of the original values, makes the method robust to models' *bias*.

The score used to assess the skill of the obtained predictions is the Roc Skill Area (RSA). The RSA takes values in $[-1, 1]$. An RSA equal to 0 indicates no skill with respect to the climatology, whilst an RSA of 1 indicates a perfect forecast [see *Jolliffe and Stephenson*, 2003, for further information on this score]. The statistical significance of the RSA score is obtained by bootstrapping [*Mason and Graham*, 2002] with 1000 samples.

For the sake of comparison, some of the figures shown trough this deliverable are related to the analysis of terciles, computed in a similar way.

4. Multi-Model Construction

In order to build the probabilities of the multi-model quintiles (or terciles) from the 45 available members (9 realizations $\times$ 5 models), we have three different options: 1) computing the quintiles for the combined series of members and models, 2) computing the quintiles separately for each model, combining the 9 available members, and 3) computing the quintiles separately for each model member. This would lead to a total of 1, 5, or 45 sets of quintiles, respectively. In order to determine which of these options was the optimum one, we computed intra-model (members) and inter-model quintile overlapping by applying an *ANOVA* test to determine whether significantly differences appear among the quintiles computed from different models (5) or model members (45). The maps shown in Figs. 1 and 2 show the *p-values* (probability of rejecting the null hypothesis of no mixing) for precipitation and maximum temperature, respectively, derived from the tests in each grid point. These figures show that the terciles/quintiles of the different members for a particular model do not overlap, whereas the terciles/quintiles of the combined multi-model ensemble overlap in several regions (overlapping mean "quintile mixing"; i.e. it indicates that the first quintile of one model may be larger than the second tercile of a different model).

As a result of this analysis we show that, for both variables, the differences among members (for a particular model) were non-significant and there was no cuantile mixing; or in other words, the nine members can be merged into a single time series (simply by averaging them), obtaining a unique set of cuantiles for all of them. However, there is a clear inter-model overlapping, indicated by the black patches in Figs. 1 and 2. Thus, cuantiles should be computed separately for each model. This inter-model overlapping is clearly larger for maximum temperature than for precipitation, indicating that the *bias* between models is larger in the former variable.

5. Results for Precipitation

5.1. Models Bias

Fig. 3 (left column) shows, by seasons, the observed VASClimO climatology and the models *bias* with respect to it at one month lead-time.

Although all models exhibit similar spatial pattern in all seasons, differences among them are visually appreciable. In general, they show a deficit of precipitation in the rainiest tropical regions of South America, Africa and Asia, but an overestimation in the extra-tropics. *Bias* are large on the whole, being frequent, for instance, values of 1000 (-1000) mm/season in regions where it rains 1000 (6000) mm/season.

In DJF, all models tend to underestimate in the rainiest parts of South America, whilst in general they overestimate in the driest parts of North America, Europe and south-eastern Asia. In Africa, all models except the one from MF underestimate in the south-eastern part of the continent (Kenya, Tanzania, Mozambique, Uganda, Congo, Zambia, Malawi and Zimbabwe), the wettest one. However, all tend to overestimate in the south and south-western, where rains are not so abundant.

In general, the spatial pattern of all models is also similar in MAM, underestimating in northern South America and overestimating, on the whole, in the northern hemisphere. In Africa, all except the MF one underestimate along the eastern coast of the continent. Models from the CMCC-INGV and the UKMO overestimate along the northern coast of the gulf of Guinea. The MF model clearly overestimates over the southern part of the continent.

Again, the spatial pattern of all models is similar in JJA. There is a common underestimation over the rainiest parts of South America and south-eastern Asia. The ECMWF, the CMCC-INGV and the UKMO models underestimate over certain parts of North America. It seems to be a little (large in the CMCC-INGV model) generalized underestimation over Europe. In Africa, all models underestimate along the northern coast of the gulf of Guinea, which is actually the rainiest region of the continent. Some of them, as the ones from the CMCC-INGV and the UKMO overestimate in neighbouring regions as well.

All models exhibit similar spatial patterns also in SON, with an underestimation in northern South America, around the gulf of Mexico and the Indochina region in south-eastern Asia. In general, all show an overestimation in extra-tropical latitudes of the northern hemisphere. In Africa, all models show a deficit of precipitation along the coast of the gulf of Guinea (where rains are quite abundant). The ECMWF, the IFM-GEOMAR and the CMCC-INGV models overestimate over a narrow belt up to this latter region. The IFM-GEOMAR, MF and the UKMO ones (especially these two last) overestimate over the central and central-southern part of the continent, where the observed regime of precipitation varies from 6000 to 0 mm/season.

VASClimO observations and models variability (standard deviation) is shown, by seasons, in the right column of Fig. 3. All models present similar spatial patterns in all seasons, exhibiting the largest variability in the rainiest regions (as actually observed), but differences among them are significant. In general, all exhibit lower than observed variability. Unfortunately, none is able to reproduce the observed variability in the tropical parts of South America and Africa, as well as in Europe as a whole, which reflects a models limitation. Contrarily, the region where models better reproduce the observed variability (independently of the season) is the Malay archipelago. As expected, given the smooth inherent to its construction process, the multi-model presents the lowest variability.

The largest differences between models take place for South America and Africa.

In Africa, the models capable to reproduce the largest variability are the ones from MF and the UKMO.

Fig. 4 is the analogue to Fig. 3 but for the four months lead-time predictions. Previous comments regarding the models *bias* also apply for this longer lead-time, which points out the predictions robustness. In terms of models variability, the only noticeable difference between both lead-times is that variability decreases at four months lead-time. This is an interesting result since suggests the existence of a limit lead-time beyond what predictions might not be considered useful.

5.2. Global Skill

Figs. 5 to 8 show the models (rows) skill for the driest (left column) and wettest (right column) quintiles at one month lead-time, by seasons. Intermediate quintiles turned out not skilful in any location and/or season, suggesting the models disability to predict 'normal' events.

Highest skill concentrates in tropical regions, where spatial pattern is quite robust for some season. However, skill in the rest of the globe is not so clear.

Globally, SON and DJF are the most skilful seasons, whilst MAM comes out the less one.

In DJF, highest skill is found in northern South America and some parts of the southern half of Africa and the Malay archipelago. Models from the ECMWF and the UKMO, as well as the multi-model, exhibit skill in the Somali peninsula for the driest quintile. Models from the ECMWF, the IFM-GEOMAR and the multi-model do it in southern Africa. There is a signal of skill for the wettest quintile to the east of the gulf of Guinea.

North-eastern Brazil and the Indochina peninsula are the most skilful regions in MAM, especially for the driest quintile. Unfortunately, there does not exist any defined spatial pattern of skill in any other region of the globe. In Africa, the model from the IFM-GEOMAR shows some signal of skill up to the north of the gulf of Guinea and in the south-western part of the continent for the driest quintile.

In JJA, the highest skill is found in Central America, northern Brazil, the gulf of Guinea and the Malay archipelago. In the gulf of Guinea, models from the IFM-GEOMAR, the CMCC-INGV, MF and the UKMO, as well as the multi-model, show some skill for the driest quintile. For the wettest one, skill patches spread over South America and the Malay archipelago, covering eastern Australia in some cases. Models from the CMCC-INGV, MF and the UKMO, as well as the multi-model, still exhibit a very little signal of skill over the gulf of Guinea.

The highest skill in SON concentrates over northern South America, a belt in Central Africa, parts of Middle East, the Malay archipelago and Australia. Signal intensity is doubtless higher in this season than in any other. The spatial pattern of skill is quite well-defined in all models, both for the driest and wettest quintiles. For the driest one, there is a large agreement between models, showing high skill in northern South America, the Malay Archipelago and eastern Australia. Models from the ECMWF, MF, the UKMO, as well as the multi-model, do the same in Arabia and Near East. All models exhibit some skill over the central part of Africa. The ones from the ECMWF and MF, as well as the multi-model, show skill in the Somali peninsula. For the wettest quintile, the spatial pattern of skill is still intense and consistent between models in northern South America, the Malay archipelago and Australia. The Arabian signal remains practically unaltered, whilst the African one shifts to the eastern part of the continent. All models present skill over a relatively vast region close the Somali peninsula. The less skilful model for Africa is the one from the CMCC-INGV.

One can see from the preceding comments that skilful regions are basically the same for the driest and the wettest quintiles in all models and seasons. To explain this, one might take into account, for instance, that a large RSA for the driest quintile does not necessarily imply skill in forecasting dry conditions, but it may be the result of forecasting non-dry conditions (any of the other four quintiles).

Fig. 9 shows the multi-model skill for the driest and wettest quintiles (columns) by seasons (rows) at four months lead-time. As can be seen, skill at four months lead-time is in general lower than at one month-lead time (which is not surprising). At this longer lead-time, skill patches only survive in regions that were highly skilful at one month lead-time. For instance, the signal of skill that appears near the Somali peninsula in SON at one month lead-time disappears at four months lead-time. The little skill obtained in MAM (the lees skilful season at one month lead-time) disappears completely at four months lead-time. SON and DJF are still the most skilful seasons at this longer lead-time.

Figs. 10 and 11 show the multi-model skill for the driest and wettest terciles (columns) by seasons (rows) at one and four months lead-time, respectively. In terms of skill, all the features found in the analysis of quintiles remain valid for the terciles. The essential difference is that skill is higher in the case of quintiles, which means that models are able to satisfactorily predict finer intervals.

6. Results for Maximum Temperature

6.1. Models Bias

As Fig. 3, Fig. 12 (left column) shows, by seasons, the observed CRU climatology and the models *bias* with respect to it at one month lead-time.

Contrarily to what occurs for precipitation, models do not exhibit common similar spatial patterns and they are very smooth. On the whole, all models underestimate maximum temperature in the major part of the globe, except in the northern latitudes of the northern hemisphere (overall the ECMWF and MF models in DJF) and in South America in some cases. In general, the largest similarity between models take place, on the one hand, for the ECMWF and MF ones, and on the other hand, for the IFM-GEOMAR and the UKMO ones. The CMCC-INGV model clearly show the largest (negative) *bias* all around the globe. All underestimate in Greenland, which probably has to do with the lack and bad quality of observations in this region.

In DJF, the ECMWF and MF models (especially the latter) overestimate in northern Asia (Russia) and northern North America, the coldest regions. This overestimation is clear in Siberia, where the rest of models also overestimate. In general, all models (especially the one from the CMCC-INGV) underestimate in the rest of the globe. In Africa, all models underestimate in the northern and southern parts of the continent.

In MAM, models from the ECMWF and MF exhibit a very light overestimation over northern Asia and northern North America. The IFM-GEOMAR overestimates slightly over southern Asia. All models except the CMCC-INGV one (which clearly underestimate in the entire globe) overestimate slightly in the south of Arabia. In general, all underestimate in the rest of the globe. *Bias* in Africa are not large, except for the MF model, which underestimates in the southern part of the continent, and obviously for the CMCC-INGV one.

In JJA, the UKMO model overestimates in general in the northern hemisphere, whilst the rest (except the CMCC-INGV one) only overestimate and slightly in southern Arabia and points of southern Asia. The MF model also overestimate in a region of northern South America. All underestimate in the rest of the globe. The only model that presents large *bias* in Africa is the CMCC-INGV one, that clearly underestimates.

In SON, the MF model overestimates in Siberia, Alaska and also slightly in north-eastern South America. The IFM-GEOMAR and the UKMO ones also overestimate slightly in the latter location. All models underestimate in the rest of the globe, especially the ECMWF and the UKMO ones in the Indian region and along the Andes mountains. Again, the CMCC-INGV model clearly underestimates in the whole globe. *Bias* in Africa are not large; all models (except the CMCC-INGV one) underestimate very slightly.

CRU observations and models variability (standard deviation) is shown, by seasons, in the right column of Fig. 12. The observed largest variability takes place overall for the coldest regions. In general, all models tend to exhibit (although to a different extent) large variability in regions with the largest observed variability (basically the northern hemisphere), but insufficiently. Unfortunately, as for precipitation, models variability is overall clearly lower than the observed. In general, contrarily to what occurs for precipitation, not all models present similar spatial patterns.

The IFM-GEOMAR and MF models present higher than observed variability in South America in all seasons, especially in DJF. The ECMWF and the UKMO do the same in South America in SON. The UKMO one exhibit higher than observed variability in the Himalayas mountains in DJF and MAM. In Africa, the largest variability is observed in the southern part of the continent in DJF and MAM, and only the ECMWF and the UKMO models are able to reproduce this (although not sufficiently).

As expected, the multi-model variability is the lowest.

Fig. 13 is the analogue to Fig. 12 but for the four months lead-time predictions. As can be seen, *bias* do not change appreciably (neither in spatial distribution nor in magnitude) with respect to the one month lead-time case. Variability keeps similar spatial patterns at this longer lead-time but is lower in magnitude.

6.2. Global Skill. De-trended data

In the validation of maximum temperature there is a new factor that makes it more delicate than that of precipitation: trends. One must take trends into account, especially when validating models, since they could lead to artificial skill. For this reason, we validated maximum temperature considering both the de-trended and the trended data. This section is devoted to the first case.

Trends were calculated grid point by grid point (for observations and for every model) using a Mann-Kendall test. Those statistically significant at a confidence level of 90% were removed from the corresponding time series. Fig.14 shows the CRU and models (at one month lead-time) trends, by seasons (columns). Points on the maps mark trends statistically significant at a 90% confidence level, that is, those that were removed. We do not want to go into detail in the analysis of trends itself since the objective in this deliverable is rather to assess their effects on skill, by qualitatively comparing the skill obtained with the de-trended data with that obtained with the trended ones. The important conclusion to be extracted from Fig. 14 is that trends are statistically significant over vast regions of the globe in all seasons.

Figs. 15 to 18 show, by seasons, the models (rows) skill for the coldest (left column) and hottest (right column) quintiles at one month lead-time, when removing statistically significant trends both in observations and models. As for precipitation, no skill was found for any of the intermediate quintiles in any location and/or season.

As for precipitation, highest skill concentrates in tropical regions, where spatial pattern is very well-defined sometimes and models agreement tend to be large. Unfortunately, this does not apply for the rest of the globe in general. Skill for maximum temperature is in general higher than for precipitation.

Globally, DJF and MAM are the most skilful seasons, whilst JJA results the less one.

In DJF, highest skill was found in the northern part of South America and the southern half of Africa, for both the coldest and hottest quintiles. There exists quite a large agreement between models over the African region. The ones showing less skill in this area are those from MF (for both coldest and hottest quintiles) and the UKMO (for the coldest quintile).

The spatial pattern of skill in MAM is similar to that of DJF but slightly less intense. Again, highest skill is located in northern South America and the southern part of Africa for both the coldest and hottest terciles. There also exists a consistent signal of skill over the coast of Guinea for the hottest quintile. For Africa, models agreement is good and the less skilful results correspond to models from MF and the UKMO.

In JJA, all models lead to lower skill than in the rest of seasons. Spatial pattern in South America continues centred over the northern part of the continent, whilst in Africa it appears shifted to the north (with respect to DJF and MAM). Although signal intensity over the northern part of Africa is not as high as in the preceding seasons, models agreement is still quite consistent, especially for the coldest quintile. The less skilful model in Africa in JJA is the one from the CMCC-INGV. It is also remarkable the fact that all models reach a considerable skill over India for the hottest quintile.

All models exhibit high skill in the northern half of South America in SON for both the coldest and hottest quintiles. Signal in Africa is probably not as spread as in JJA but it is slightly more intense and more concentrated over the central part of the continent. As in JJA, the spatial pattern in SON is not so well-defined as the obtained in DJF and MAM.

Fig. 19 shows the multi-model skill for the coldest and hottest quintiles (columns) by seasons (rows) at four months lead-time. Important results can be extracted from it. First, it is striking the fact that skill obtained at one month lead-time in MAM (the most skilful season with DJF) disappears completely at four months lead-time (the same occurs with precipitation in this season). Unfortunately, we did not find any explanation for this. Second, skill at four months lead-time is in general slightly lower than at one month lead-time (as expected) and its spatial pattern spreads, losing its definition. DJF is still the most skilful season at this longer lead-time. Furthermore, regions that are not skilful at one month lead-time appear as skilful ones at four months lead-time, such as north-western Russia in MAM (hottest quintile).

Figs. 20 and 21 show the multi-model skill for the coldest and hottest terciles (columns) by seasons (rows) at one and four months lead-time, respectively. As for precipitation, all features found in the analysis of quintiles remain valid for terciles, being the difference that skill is higher in the case of quintiles, which means that models are able to satisfactorily discriminate finer intervals.

6.3. Global Skill. Trended data

Figs. 22 to 25 show, by seasons, the models (rows) skill for the coldest (left column) and hottest (right column) quintiles at one month lead-time, when retaining trends in both observations and models. The scope here is to qualitatively analyse, overall, the main differences with respect to the de-trended validation, without going further into details. When comparing Figs. 15 to 18 (de-trended data) with Figs. 22 to 25 (trended data), one sees that spatial pattern of skill is very similar in both cases, but in general skilful regions spread and the intensity of the signal increases slightly in the trended case. It occurs specially in Africa, the only region where all models reproduce significant trends in all seasons. This suggests thus that trends effectively inflate skill artificially and should be removed before validating.

7. Conclusions

The main aim of this work was to validate accumulated precipitation and maximum temperature seasonal predictions worldwide for the period 1961-2000. To this, we have used the hindcasts from the five state-of-the-art coupled atmosphere-ocean models within the second stream of the ENSEMBLES project, at one and four months lead-time.

As a first step, we have constructed a multi-model by applying equal weights to all models and members, which is justified by the fact of having well-defined members and models overlapping.

Results from the validation show that highest skill concentrates around the tropics for both precipitation and maximum temperature; moreover, models agreement tend to be large in these regions. Unfortunately, the latter does not apply for the rest of the globe in general. For precipitation, autumn and winter are the most skilful seasons, whilst spring is the less one. For maximum temperature, the most skilful seasons are winter and spring, whilst summer results the less one. Skill is higher when applying a quintile-based validation than when applying a tercile-based one, which means that models are able to satisfactorily discriminate finer intervals. As expected, skill decreases with longer-times. We have also checked the importance of removing trends when validating maximum temperatures since they lead to artificial skill.

To complement the study, models *bias* and their variability have been assessed. In general, models underestimate precipitation in the rainiest regions of South America, Africa and Asia, whilst overestimate in the extra-tropics. For maximum temperature, there is a generalized underestimation, with the exception of the northern latitudes of the northern hemisphere. Variability reproduced by models is lower than observed for both variables; moreover, it decreases with longer lead-times.

We acknowledge the QWeCI project, funded by the European Comission's 7^{th} Framework Programme through contract 243964, for the financial support of this work; as well as the ENSEMBLES project, funded by the European Commission's 6^{th} Framework Programme through contract GOCE-CT-2003-505539, for the data provided.

References

Batte, L., and M. Deque (2011), Seasonal predictions of precipitation over Africa using coupled ocean-atmosphere general circulation models: skill of the ENSEMBLES project multimodel ensemble forecasts, *TELLUS SERIES A-DYNAMIC METEOROLOGY AND OCEANOGRAPHY*, *63*(2), 283–299, doi: 10.1111/j.1600-0870.2010.00493.x.

Díez, E., B. ORFILA, M. D. Frías, J. Fernández, A. S. Cofiño, and J. M. Gutiérrez (2011), Downscaling ecmwf seasonal precipitation forecasts in europe using the rca model, *Tellus A*, pp. no–no, doi:10.1111/j.1600-0870.2011.00523.x.

Frías, M. D., S. Herrera, A. S. Cofiño, and J. M. Gutiérrez (2010), Assessing the skill of precipitation and temperature seasonal forecasts in Spain. Windows of opportunity related to ENSO events, *Journal of Climate*, *23*, 209–212, doi: 10.1175/2009JCLI2824.1.

Halpert, M. S., and C. F. Ropelewski (1992), Surface temperature patterns associated with the Southern Oscillation, *Journal of Climate*, *5*, 577–593.

Jolliffe, I. T., and D. B. Stephenson (2003), *Forecast Verification-A Practitioner's Guide in Atmospheric Sciences*, John Wiley & Sons.

Kirtman, B., and A. Pirani (2008), WCRP position paper on seasonal prediction, *Tech. rep.*, WCRP Informal Report No. 3/2008. ICPO Publication No. 127.

Mason, J., and N. E. Graham (2002), Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation, *Quarterly Journal of the Royal Meteorological Society*, *128*, 2145–2166.

Mitchell, T. D., and P. D. Jones (2005), An improved method of constructing a database of monthly climate observations and associated high-resolution grids, *International Journal of Climatology*, *25*(6), 693–712, doi:10.1002/joc.11813.

Palmer, T. N., A. Alessandri, U. Andersen, P. Cantelaube, M. Daveyand, P. Délécluse, M. Déqué, E. Díez, F. J. Doblas-Reyes, H. Feddersen, R. Graham, S. Gualdi, J. F. Guérémy, R. Hagedorn, M. Hoshen, N. Keenlyside, M. Latif, A. Lazar, E. Maisonnave, V. Marletto, A. P. Morse, B. Orfila, P. Rofel, J. M. Terres, and M. C. Thomson (2004), Development of a European Multimodel Ensemble system for seasonal-to-inTERannual prediction DEMETER, *Bulletin of the American Meteorological Society*, *85*, 853–872.

Schneider, U., T. Fuchs, A. Eyer-Christoffer, and B. Rudolf (2008), Global Precipitation Analysis Products of the GPCC, *Internet Publication. Updated version of Rudolf, B. (2005).*

Weisheimer, A., F. J. Doblas-Reyes, T. N. Palmer, A. Alessandri, A. Arribas, M. Déqué, N. Keenlyside, M. MacVean, A. Navarra, and P. Rogel (2009), ENSEMBLES: A new multi-model ensemble for seasonal-to-annual prediction. Skill and progress beyond DEMETER in forecasting tropical pacific SSTs, *Geophysical Research Letters*, *36*, doi: 10.1029/2009GL040896.

rmanzanas@ifca.unican.es

R. Manzanas, Instituto de Física de Cantabria, CSIC-Universidad de Cantabria.

Edificio Juan Jordá, Avenida de los Castros, S/N, 39005, Santander, SPAIN.

http://www.meteo.unican.es

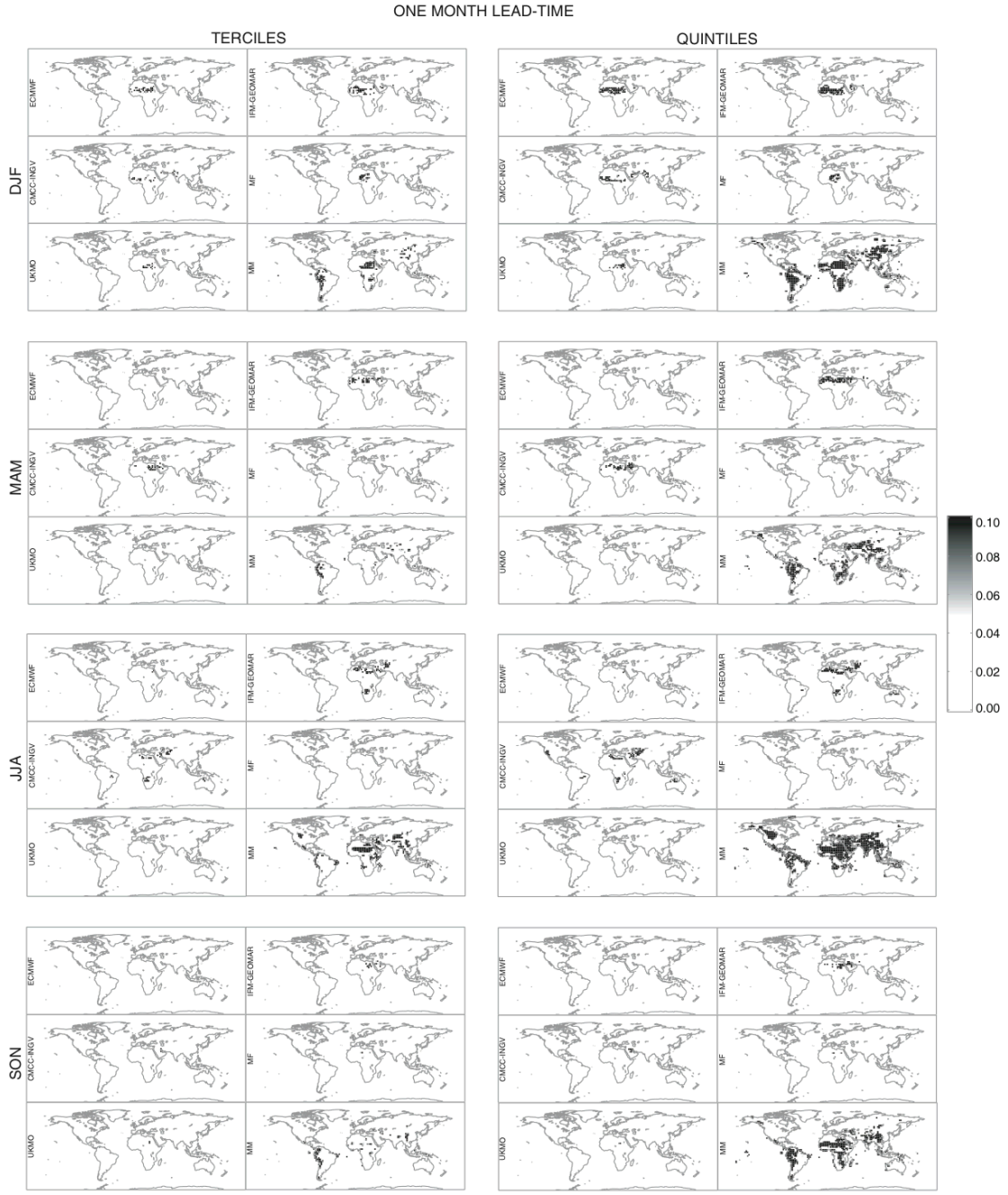

Figure 1. *P-values* for precipitation from the *ANOVA* tests performed to the terciles of their nine members of each particular model (and for the combined multi-model) for the different seasons, at one month lead-time for the terciles and quintiles. For the multi-model (MM), the *p-value* from the *ANOVA* tests is computed for the cuantiles of the 45 available model-members, at one month lead-time. Black color indicates regions where the terciles/quintiles overlap; i.e. the first quintile of one model is larger than the second quintile of a different one.

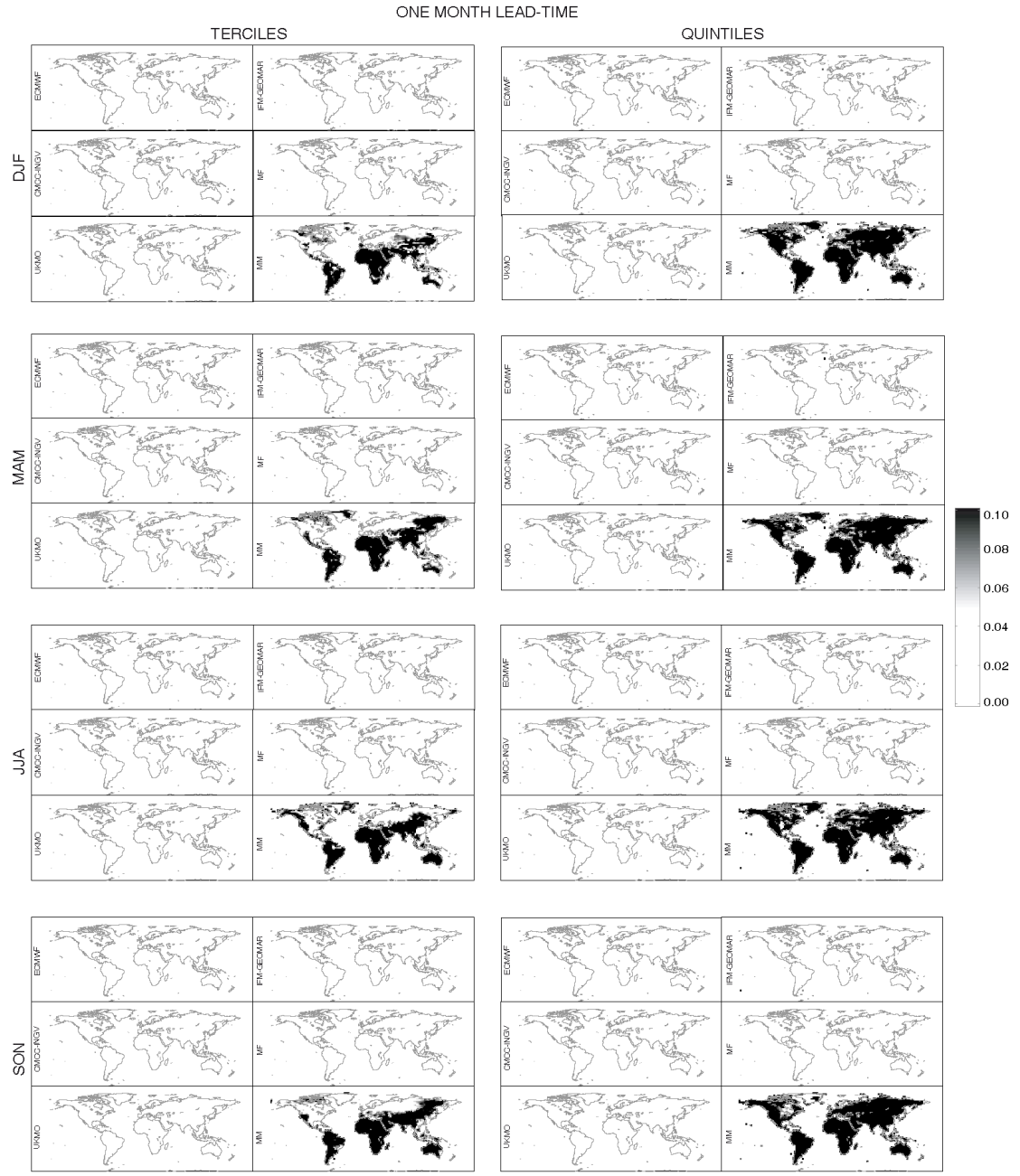

Figure 2. As Fig. 1, but for maximum temperature.

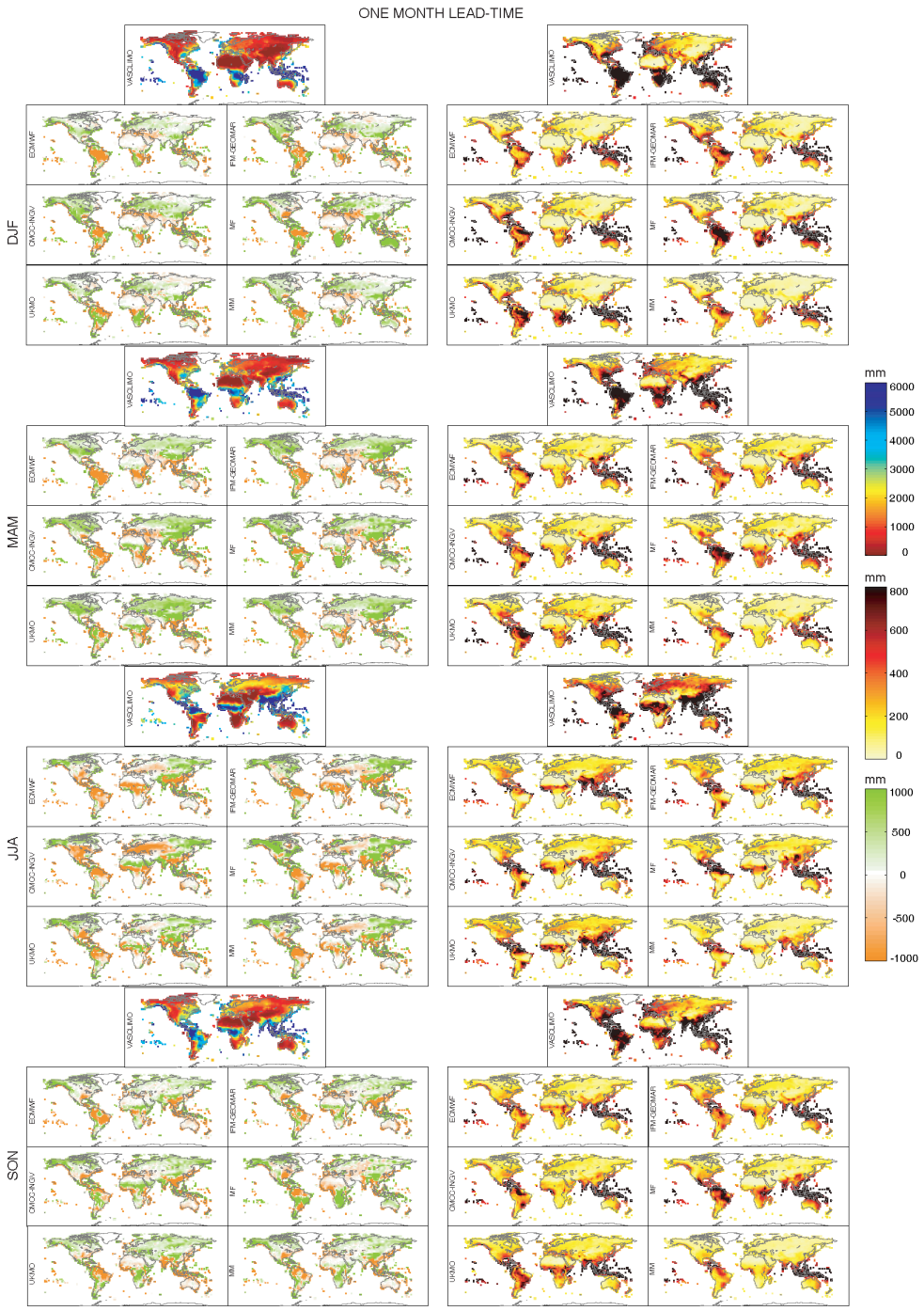

Figure 3. Left column: Observed VASClimO climatology and models *bias*, at one month lead-time (period 1961-2000), by seasons (rows). Right column: Observed VASClimO variability and models variability, at one month lead-time (period 1961-2000), by seasons (rows).

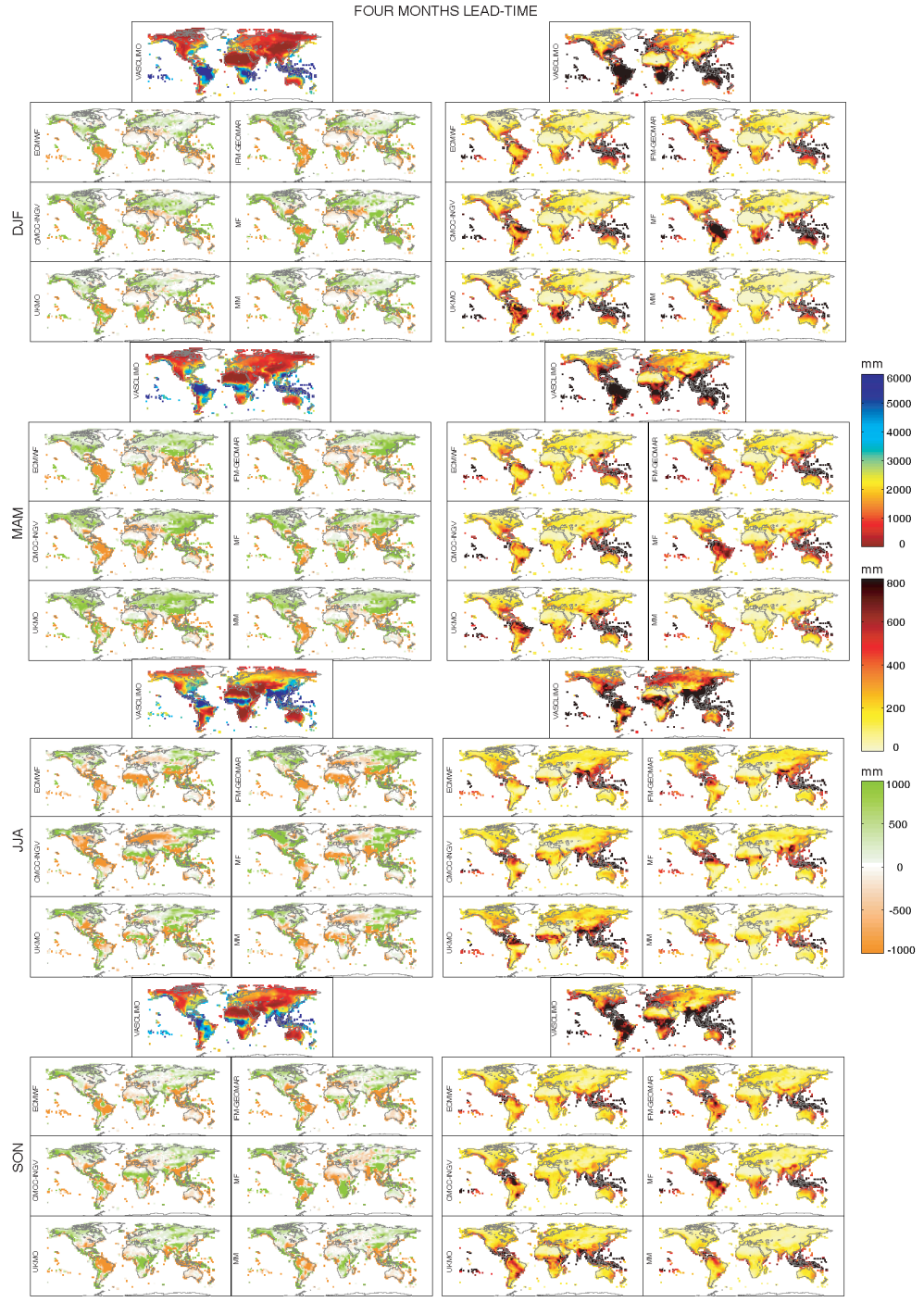

Figure 4. Left column: Observed VASClimO climatology and models *bias*, at four months lead-time (period 1961-2000), by seasons (rows). Right column: Observed VASClimO variability and models variability, at four months lead-time (period 1961-2000), by seasons (rows).

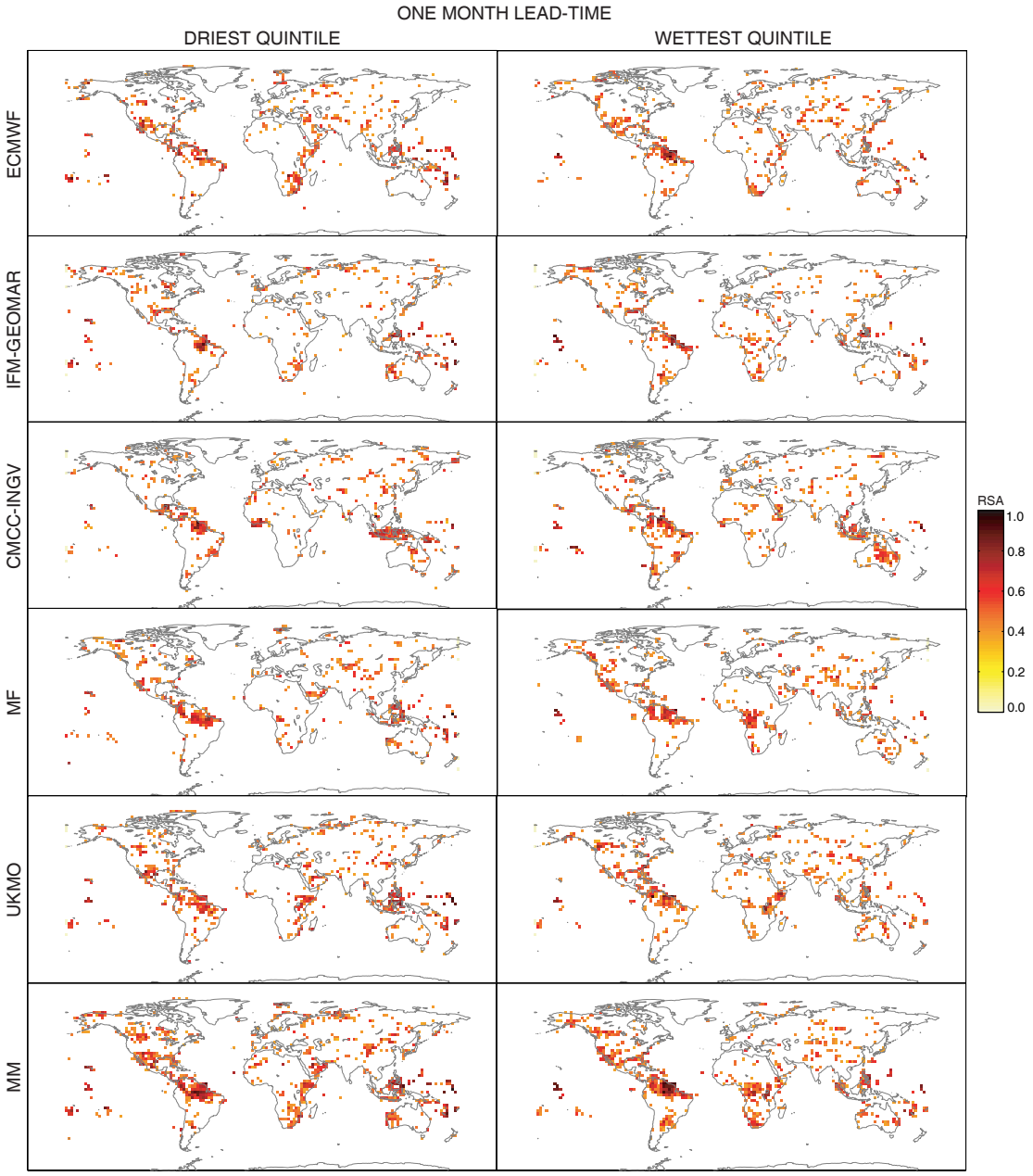

Figure 5. Left column: Models (rows) skill for the driest quintile in DJF (period 1961-2000) at one month lead-time. Right column: Idem but for the wettest quintile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

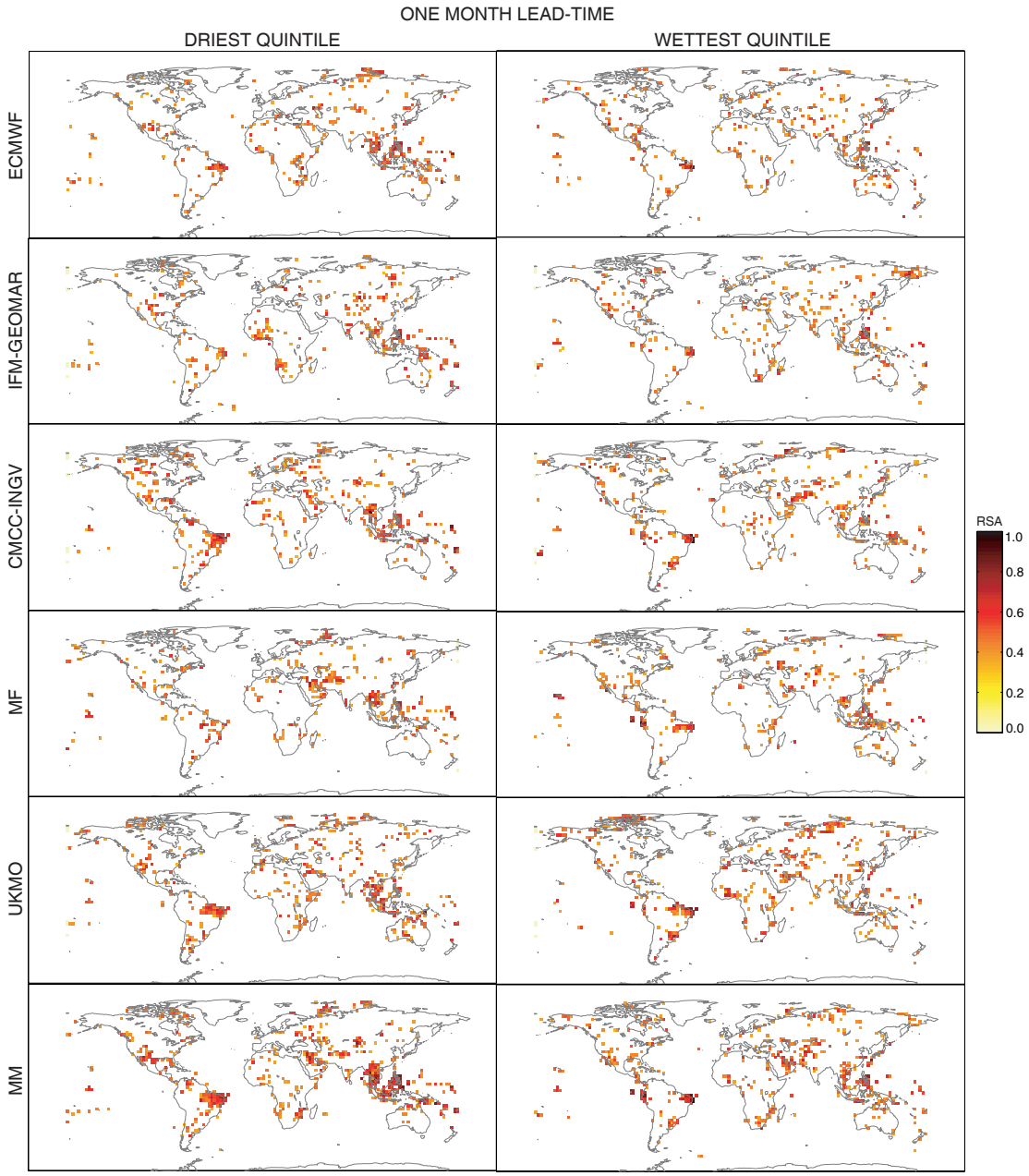

Figure 6. Left column: Models (rows) skill for the driest quintile in MAM (period 1961-2000) at one month lead-time. Right column: Idem but for the wettest quintile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

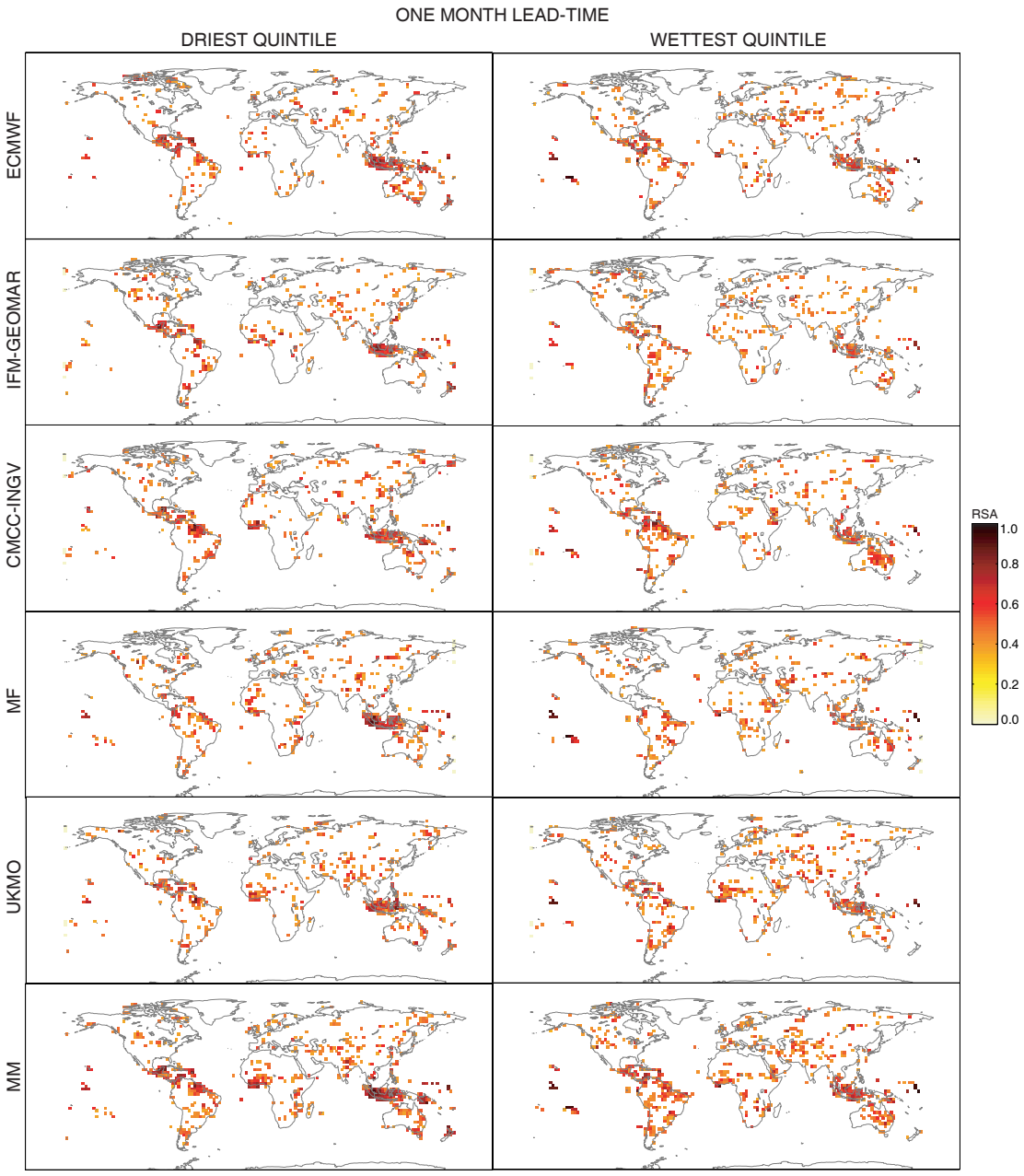

Figure 7. Left column: Models (rows) skill for the driest quintile in JJA (period 1961-2000) at one month lead-time. Right column: Idem but for the wettest quintile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

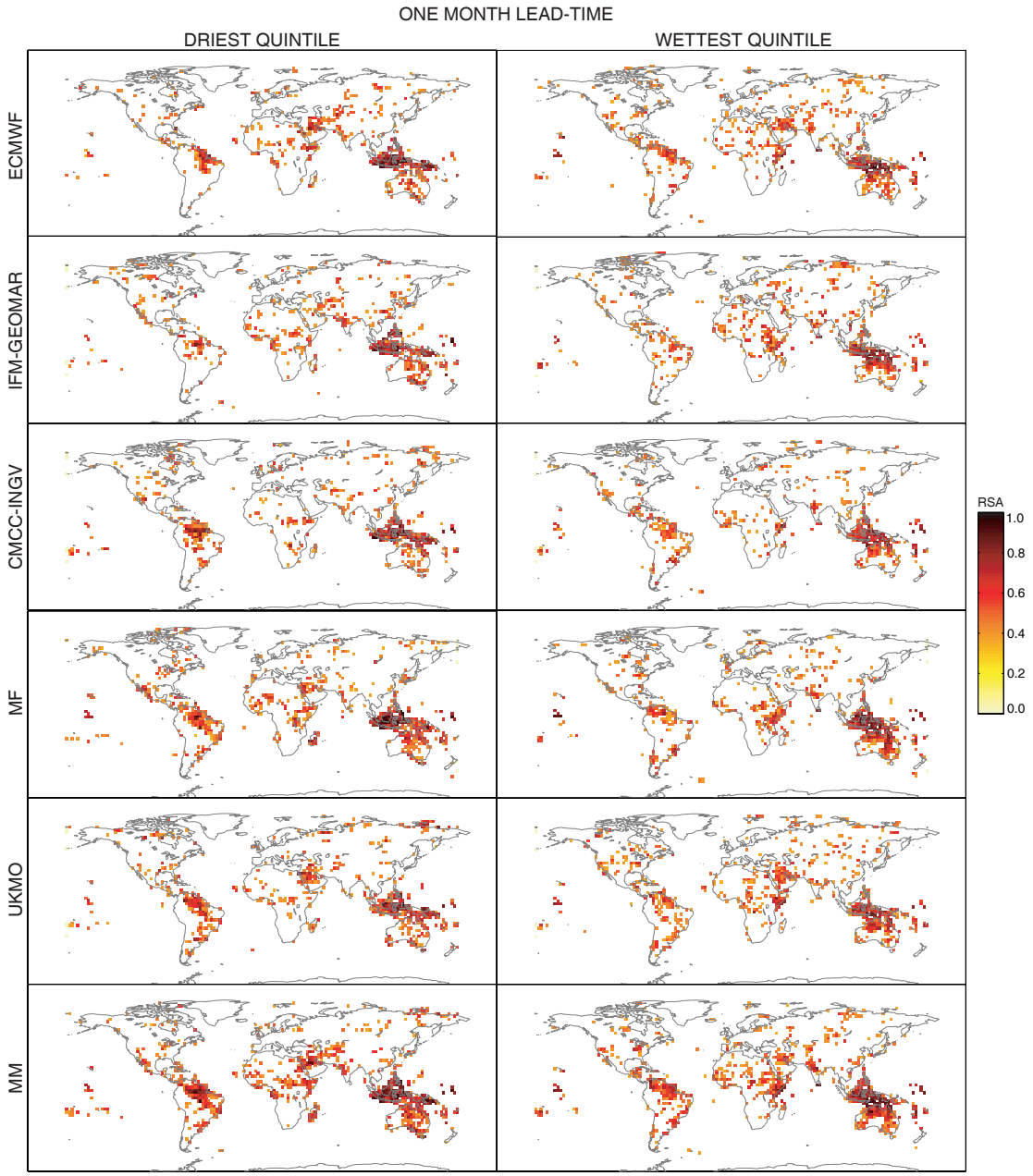

Figure 8. Left column: Models (rows) skill for the driest quintile in SON (period 1961-2000) at one month lead-time. Right column: Idem but for the wettest quintile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

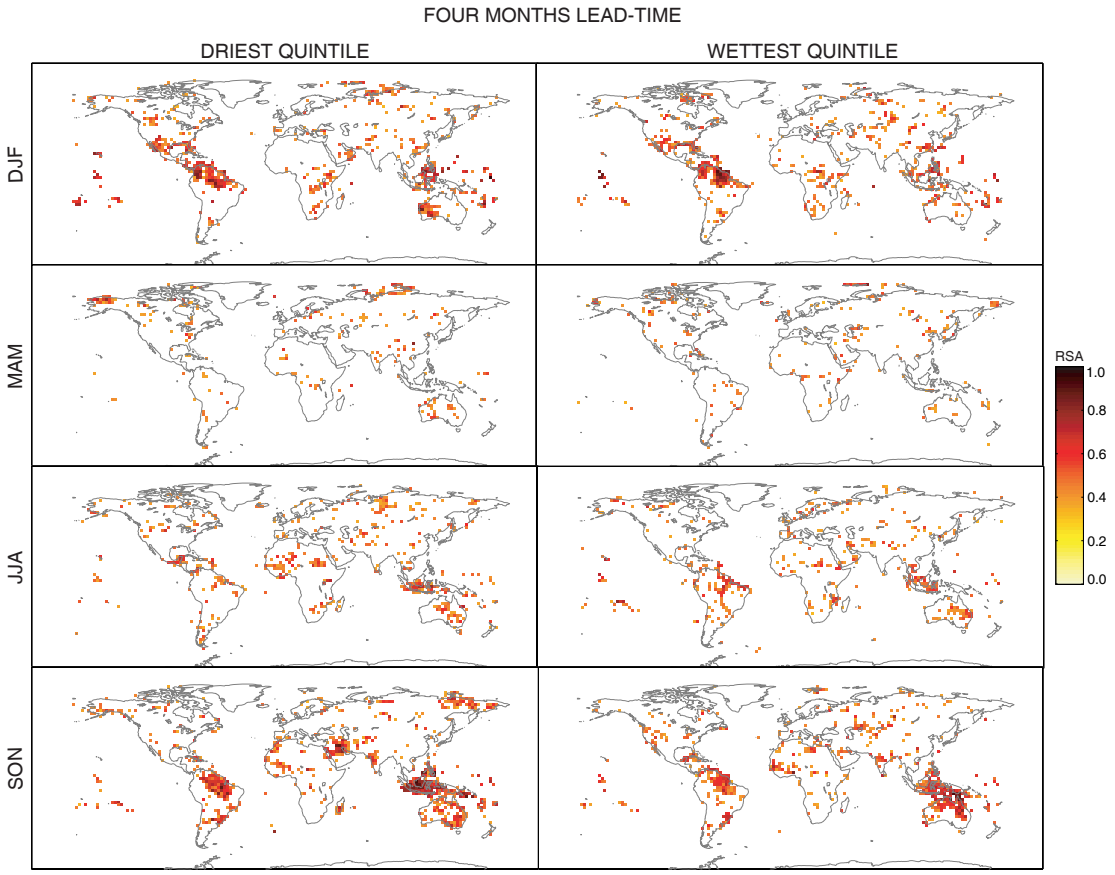

Figure 9. Left column: Multi-model skill for the driest quintile (period 1961-2000) at four months lead-time, by seasons (rows). Right column: Idem but for the wettest quintile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

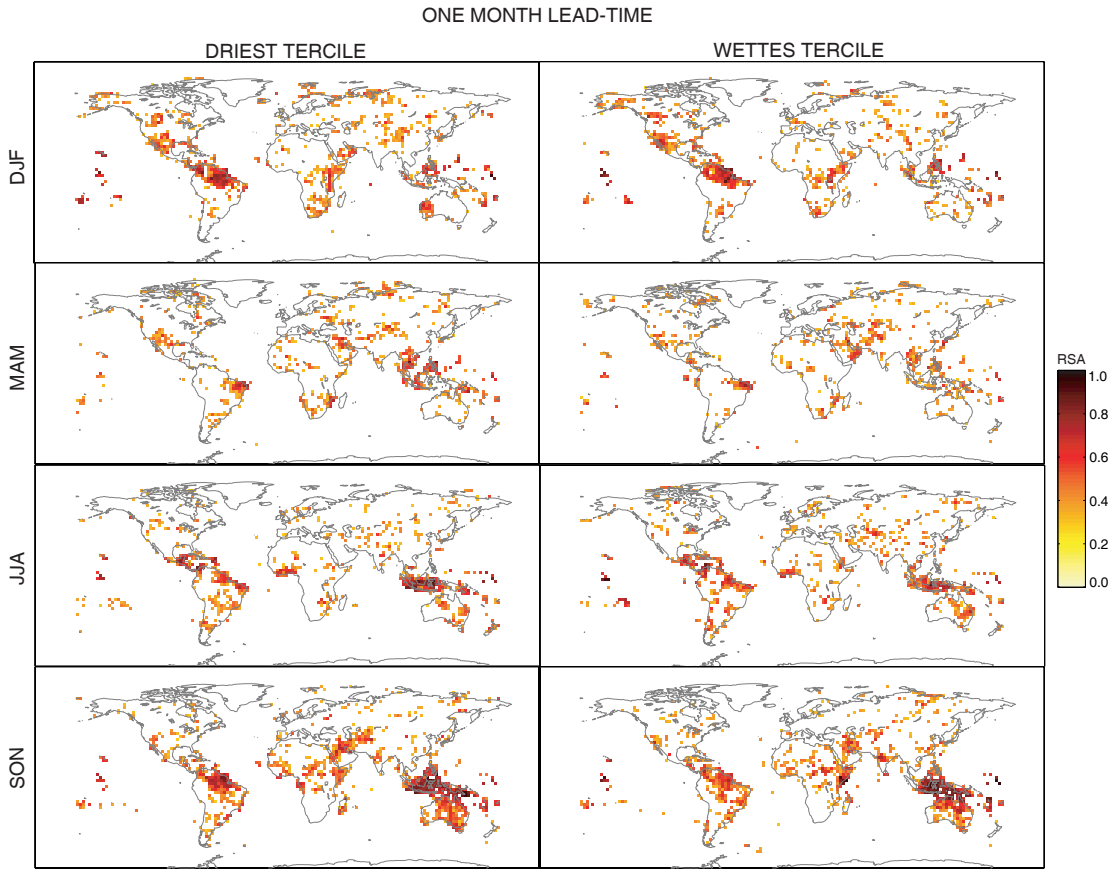

Figure 10. Left column: Multi-model skill for the driest tercile (period 1961-2000) at one month lead-time, by seasons (rows). Right column: Idem but for the wettest tercile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

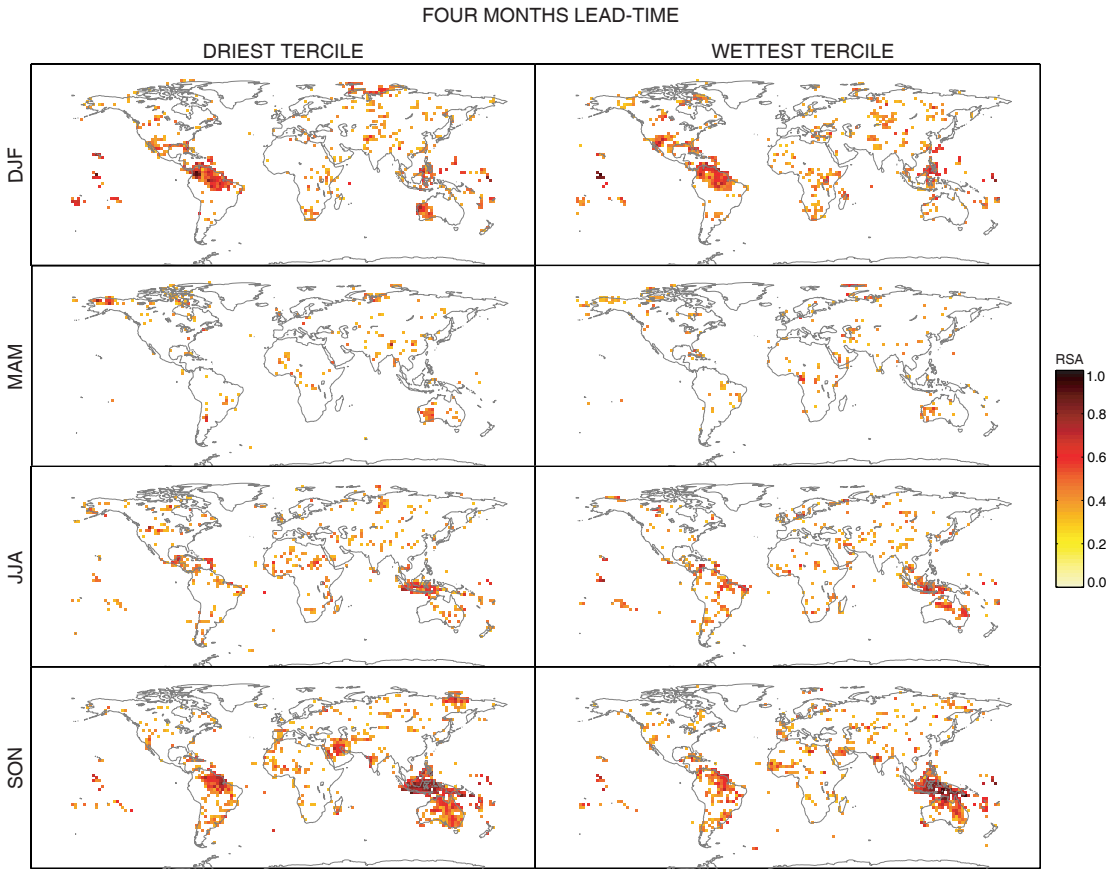

Figure 11. Left column: Multi-model skill for the driest tercile (period 1961-2000) at four months lead-time, by seasons (rows). Right column: Idem but for the wettest tercile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

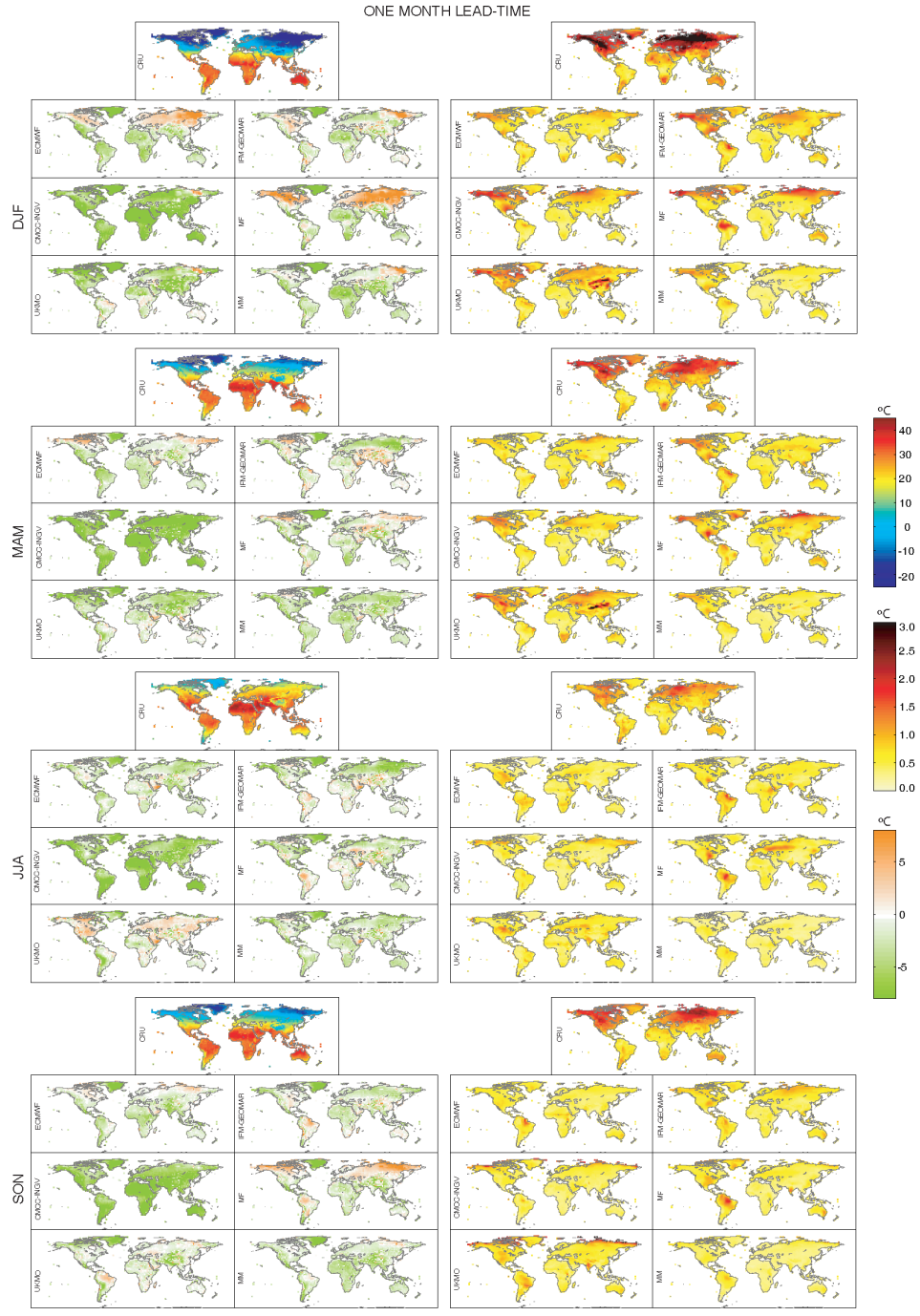

Figure 12. Left column: Observed CRU climatology and models *bias*, at one month lead-time (period 1961-2000), by seasons (rows). Right column: Observed CRU variability and models variability, at one month lead-time (period 1961-2000), by seasons (rows).

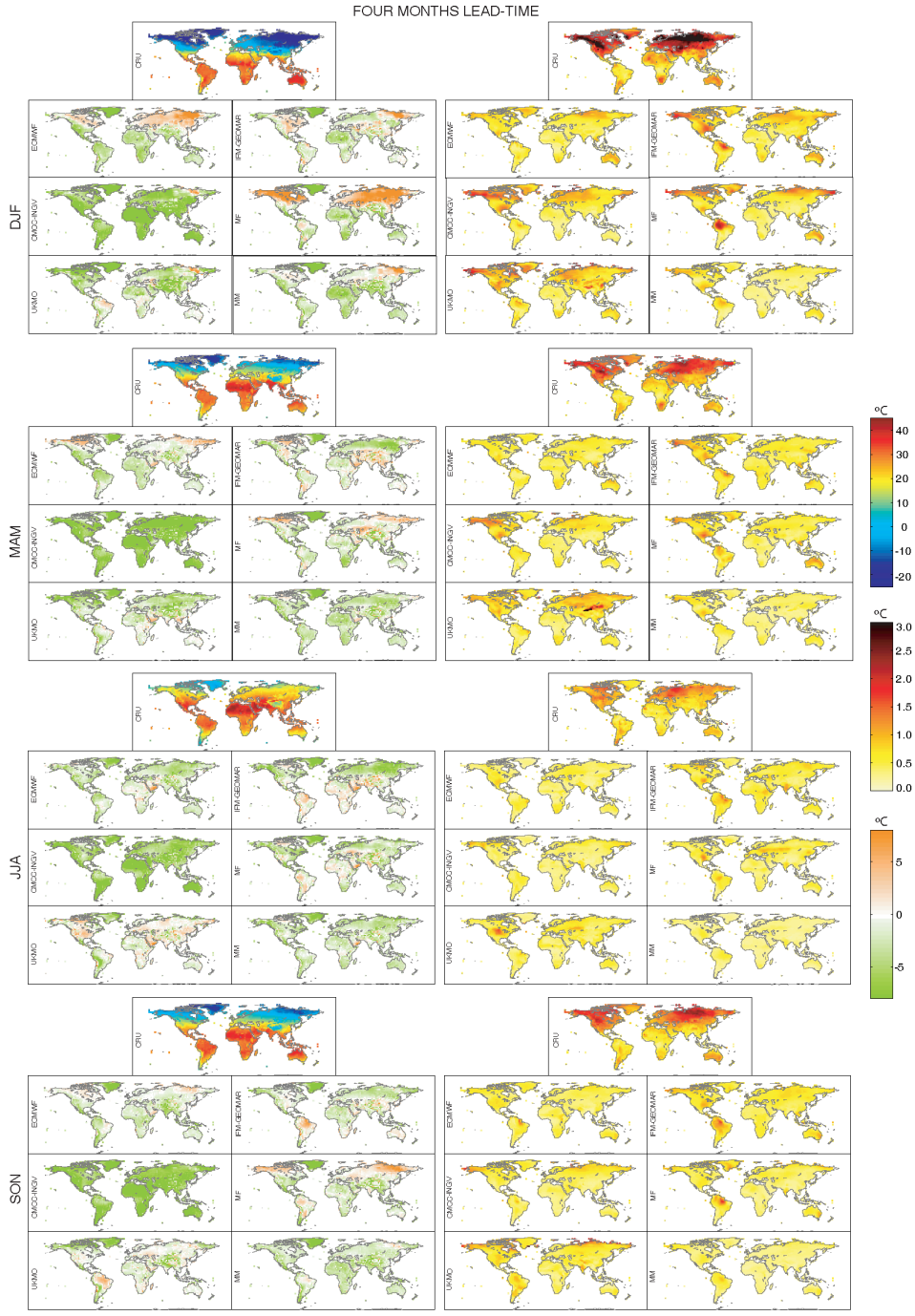

Figure 13. Left column: Observed CRU climatology and models *bias*, at four months lead-time (period 1961-2000), by seasons (rows). Right column: Observed CRU variability and models variability, at four months lead-time (period 1961-2000), by seasons (rows).

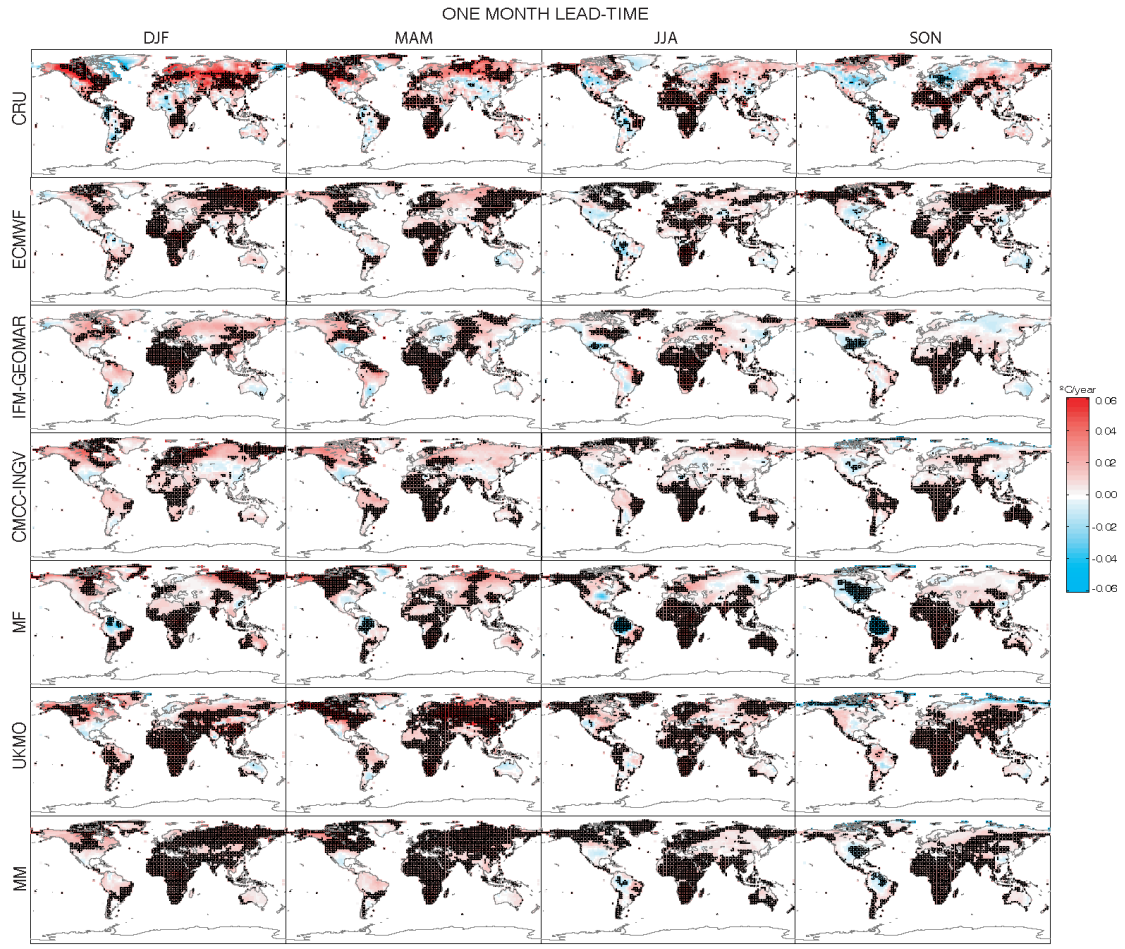

Figure 14. CRU and models (at one month lead-time) trends for the period 1961-2000, by seasons (columns). Points mark trends statistically significant at a 90% confidence level (Mann-Kendall test).

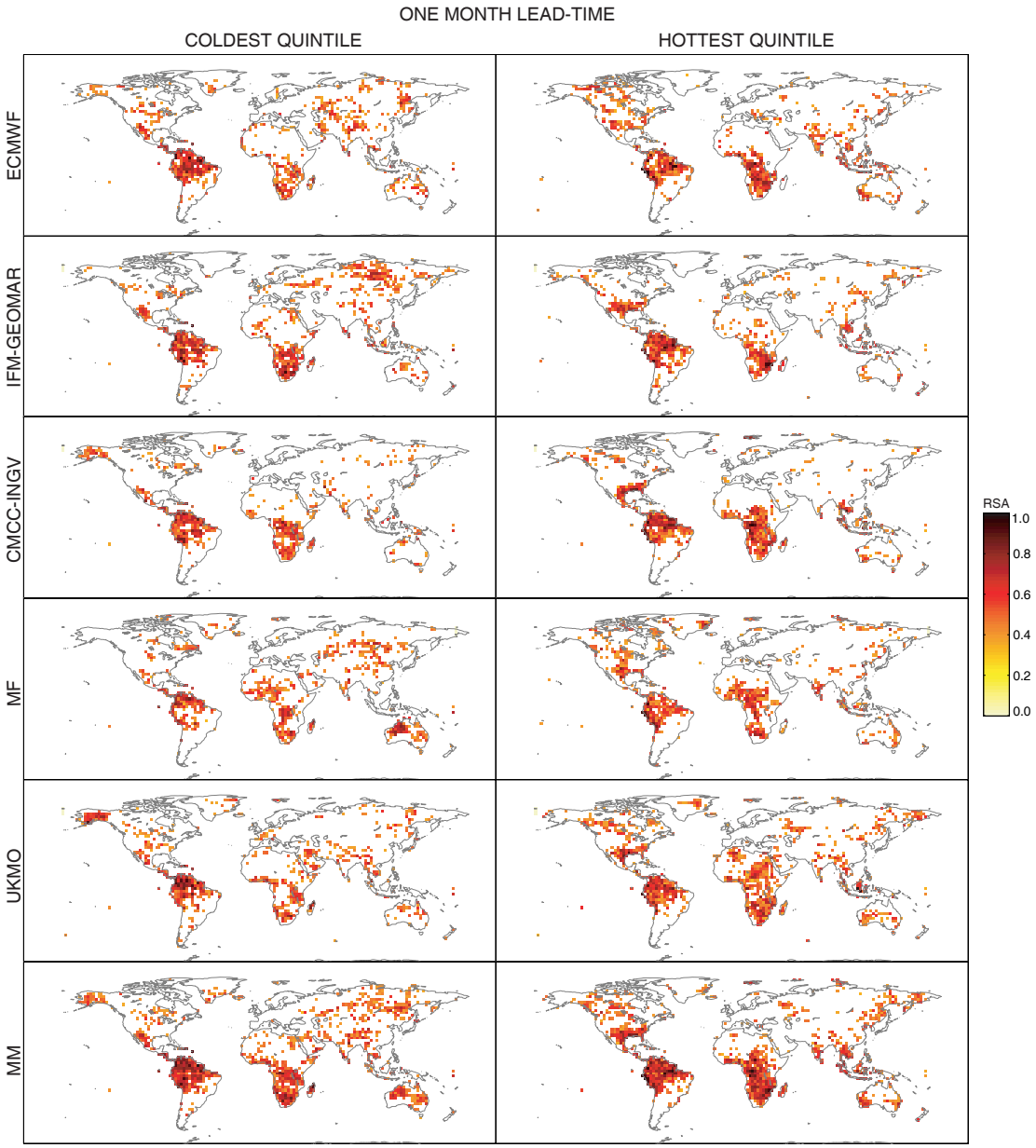

Figure 15. Left column: Models (rows) skill for the coldest quintile in DJF (period 1961-2000) at one month lead-time (de-trended data). Right column: Idem but for the hottest quintile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

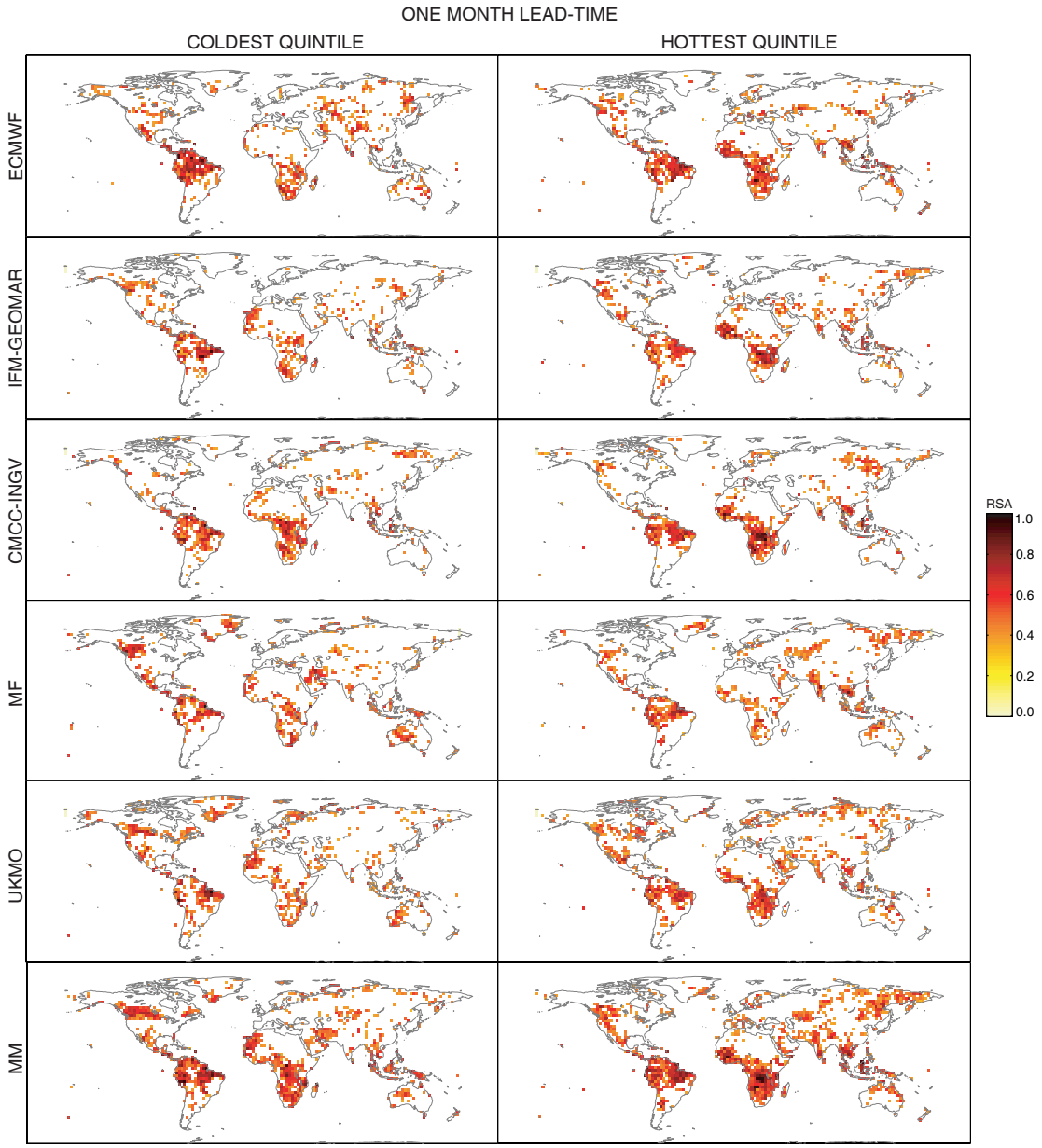

Figure 16. Left column: Models (rows) skill for the coldest quintile in MAM (period 1961-2000) at one month lead-time (de-trended data). Right column: Idem but for the hottest quintile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

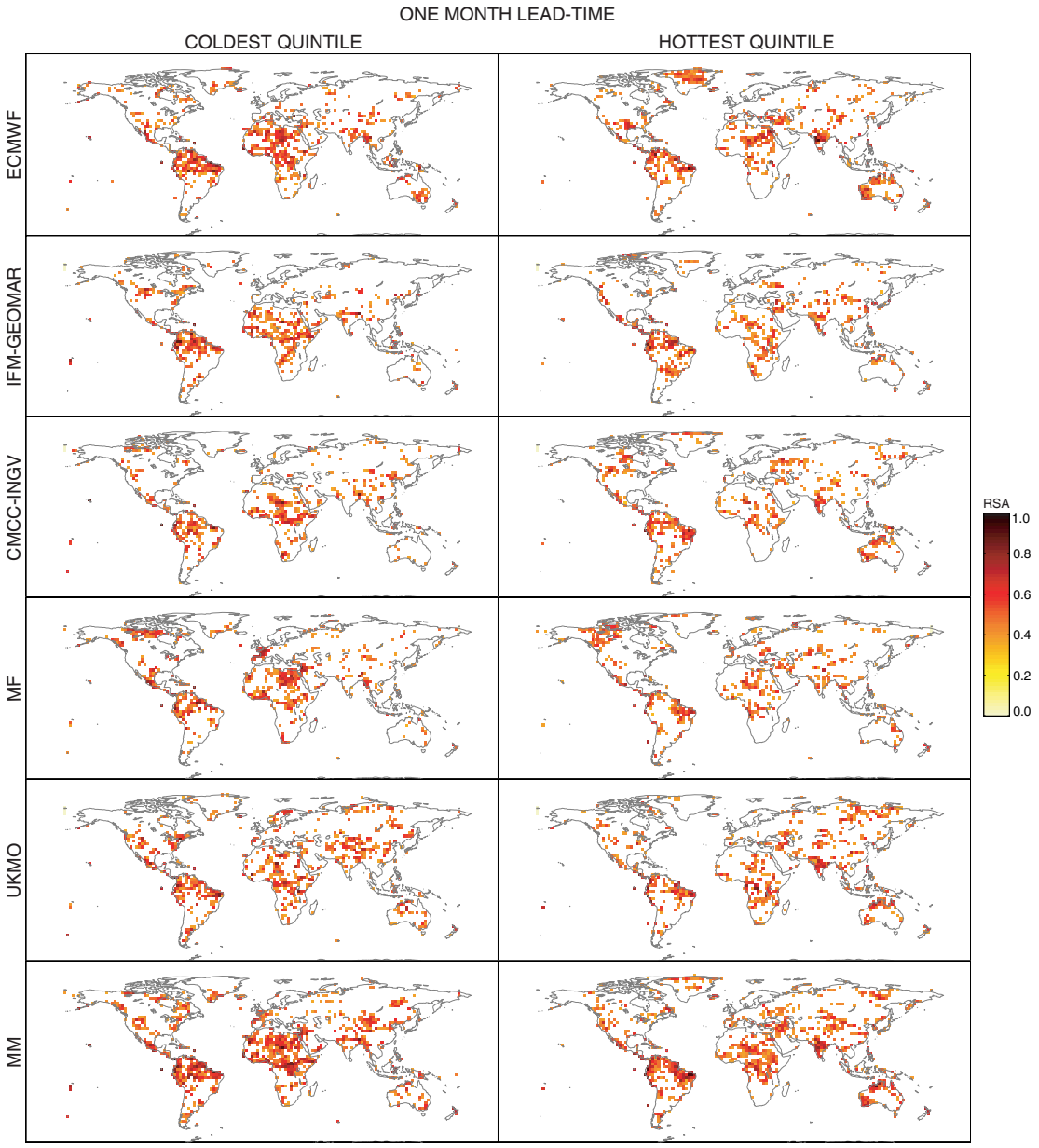

Figure 17. Left column: Models (rows) skill for the coldest quintile in JJA (period 1961-2000) at one month lead-time (de-trended data). Right column: Idem but for the hottest quintile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

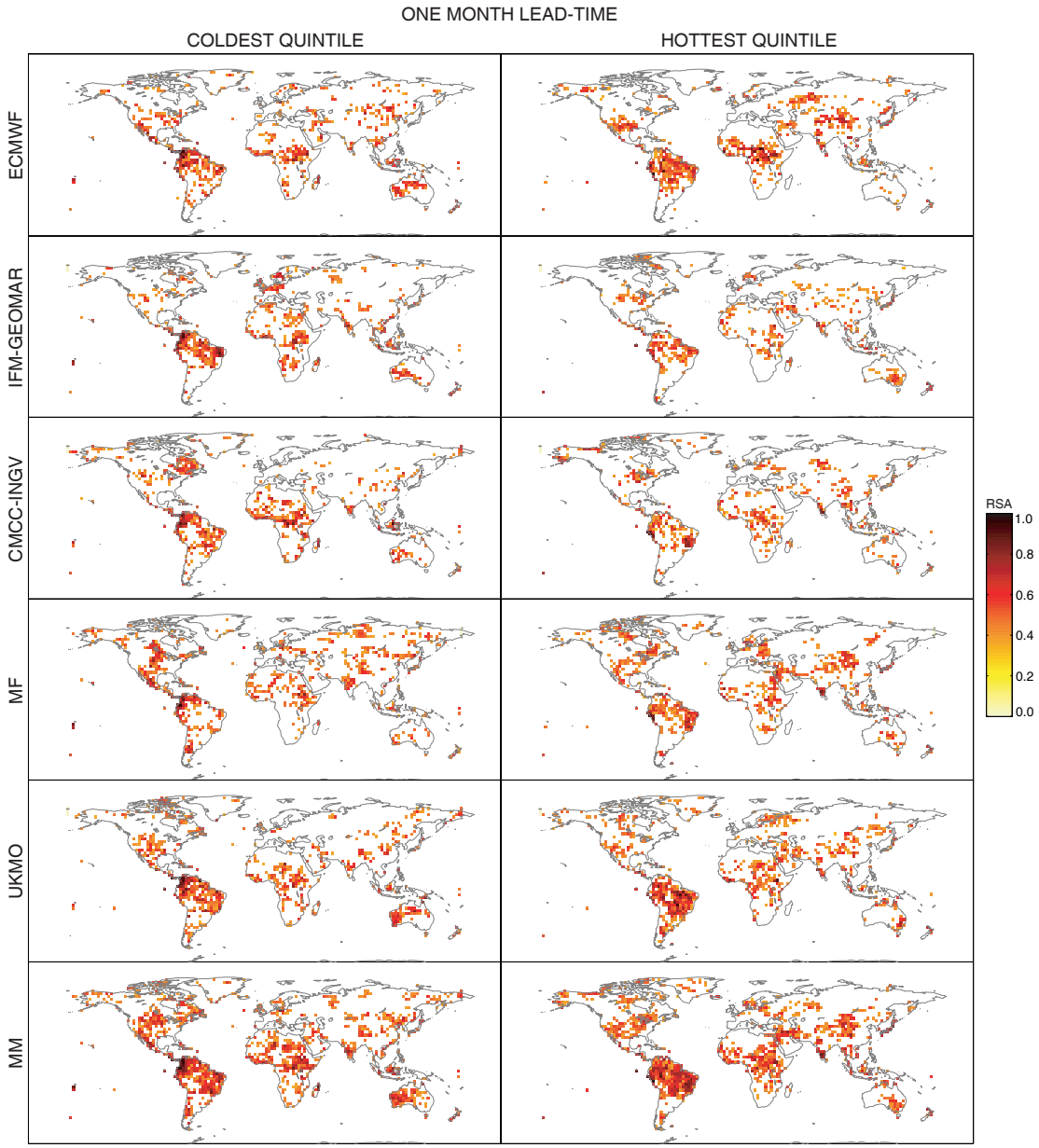

Figure 18. Left column: Models (rows) skill for the coldest quintile in SON (period 1961-2000) at one month lead-time (de-trended data). Right column: Idem but for the hottest quintile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

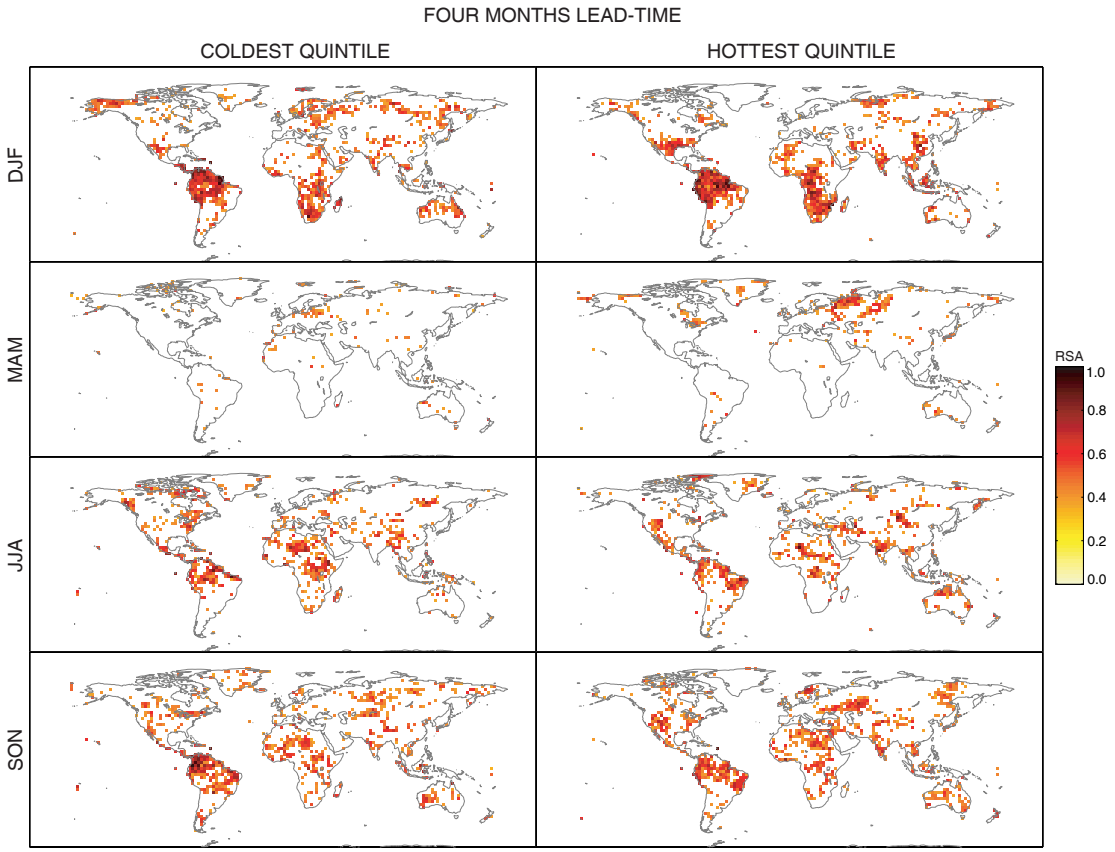

Figure 19. Left column: Multi-model skill for the coldest quintile (period 1961-2000) at four months lead-time (de-trended data), by seasons (rows). Right column: Idem but for the hottest quintile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

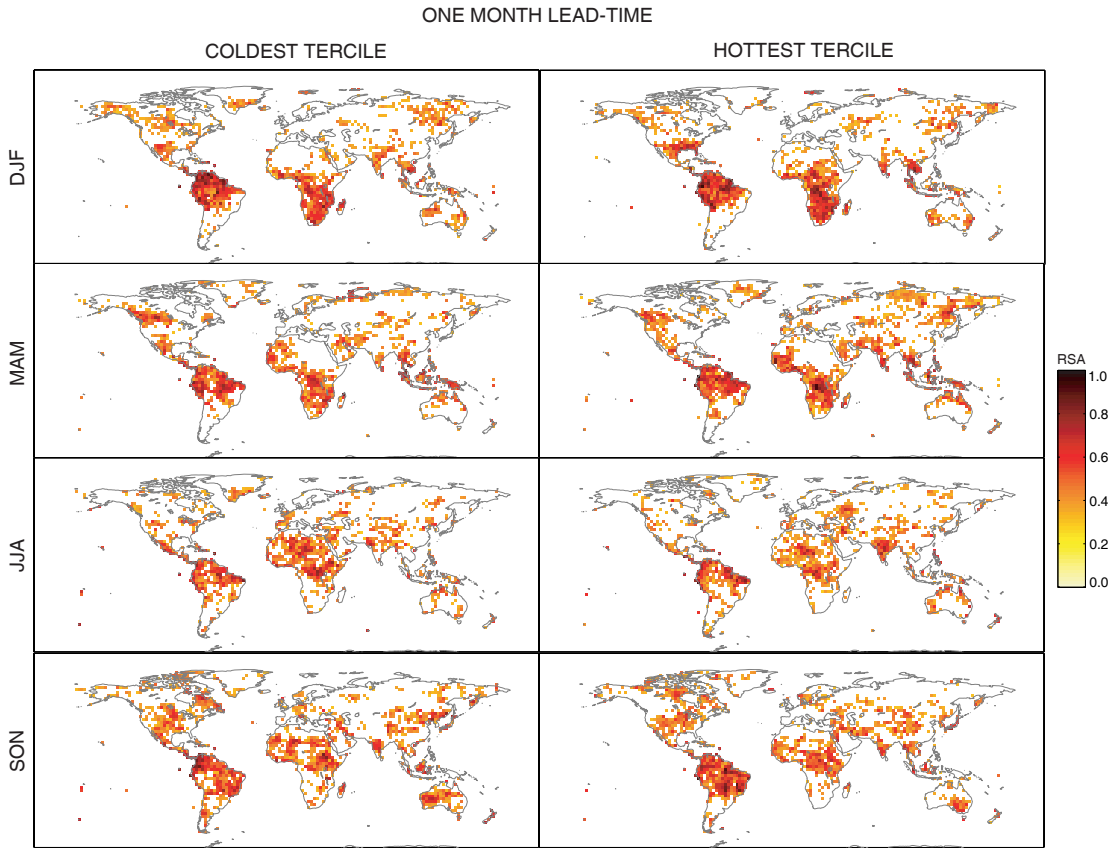

Figure 20. Left column: Multi-model skill for the coldest tercile (period 1961-2000) at one month lead-time (de-trended data), by seasons (rows). Right column: Idem but for the hottest tercile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

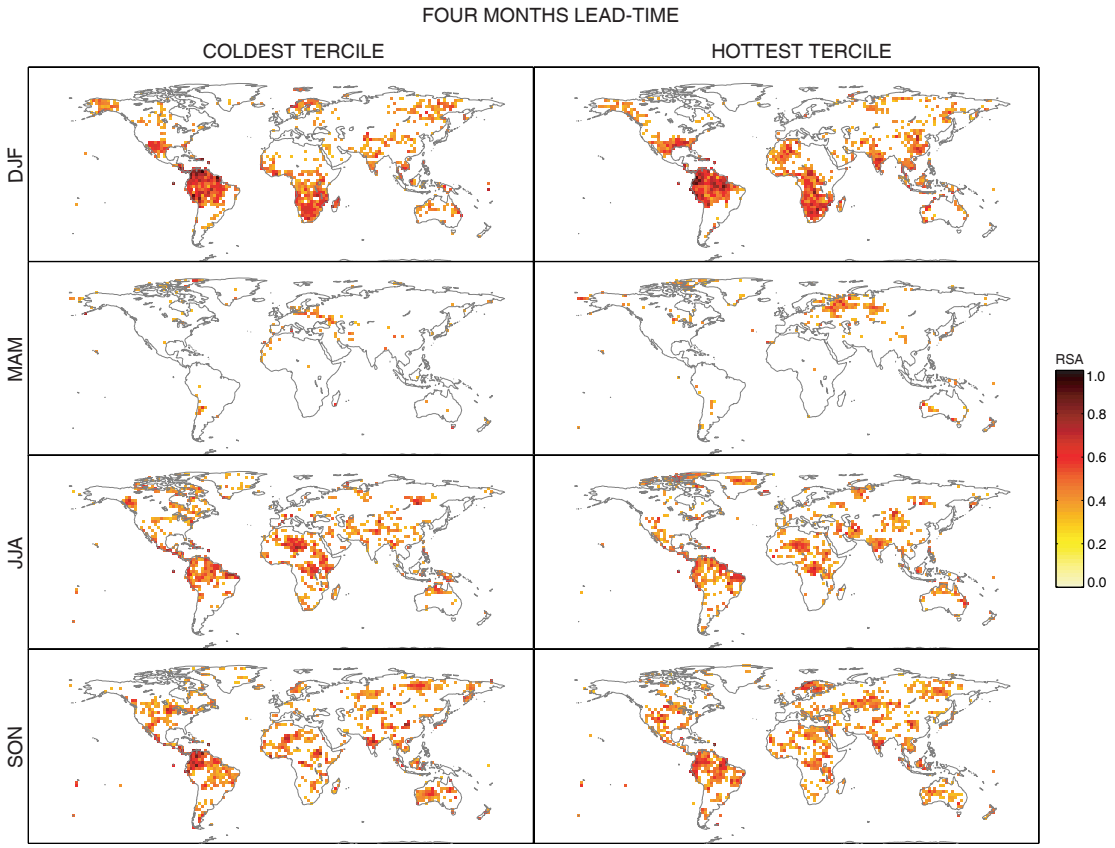

Figure 21. Left column: Multi-model skill for the coldest tercile (period 1961-2000) at four months lead-time (de-trended data), by seasons (rows). Right column: Idem but for the hottest tercile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

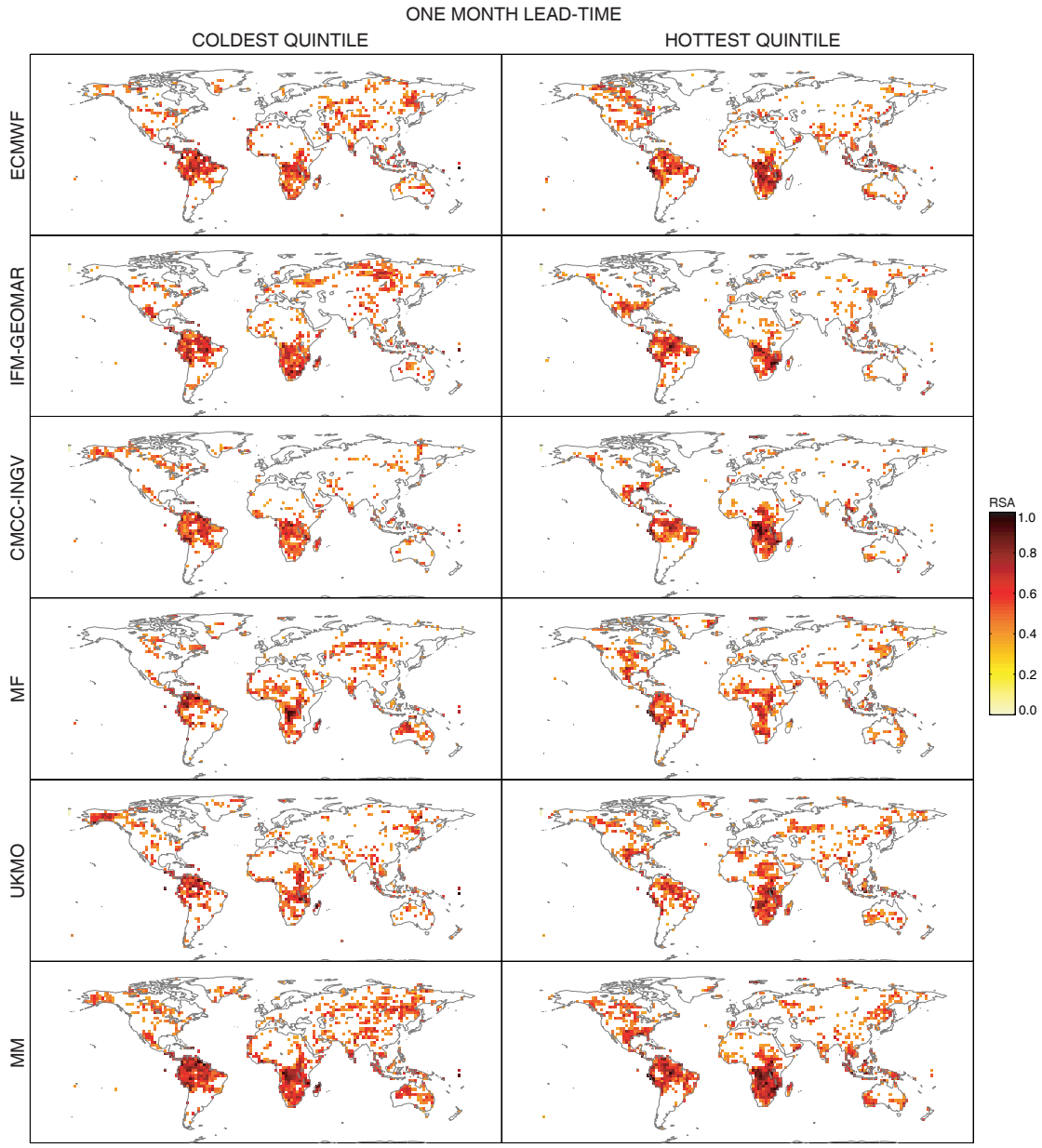

Figure 22. Left column: Models (rows) skill for the coldest quintile in DJF (period 1961-2000) at one month lead-time (trended data). Right column: Idem but for the hottest quintile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

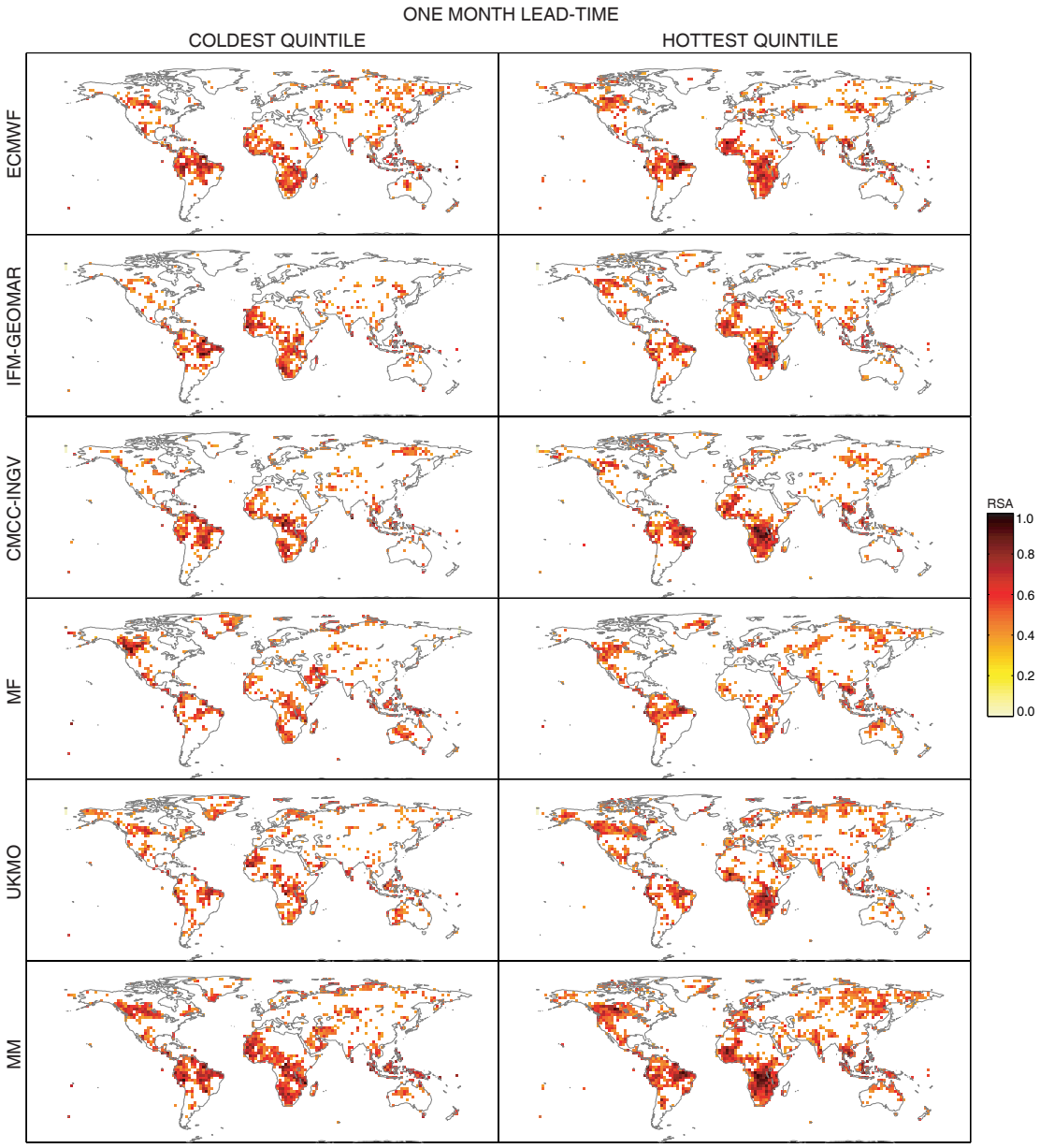

Figure 23. Left column: Models (rows) skill for the coldest quintile in MAM (period 1961-2000) at one month lead-time (trended data). Right column: Idem but for the hottest quintile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

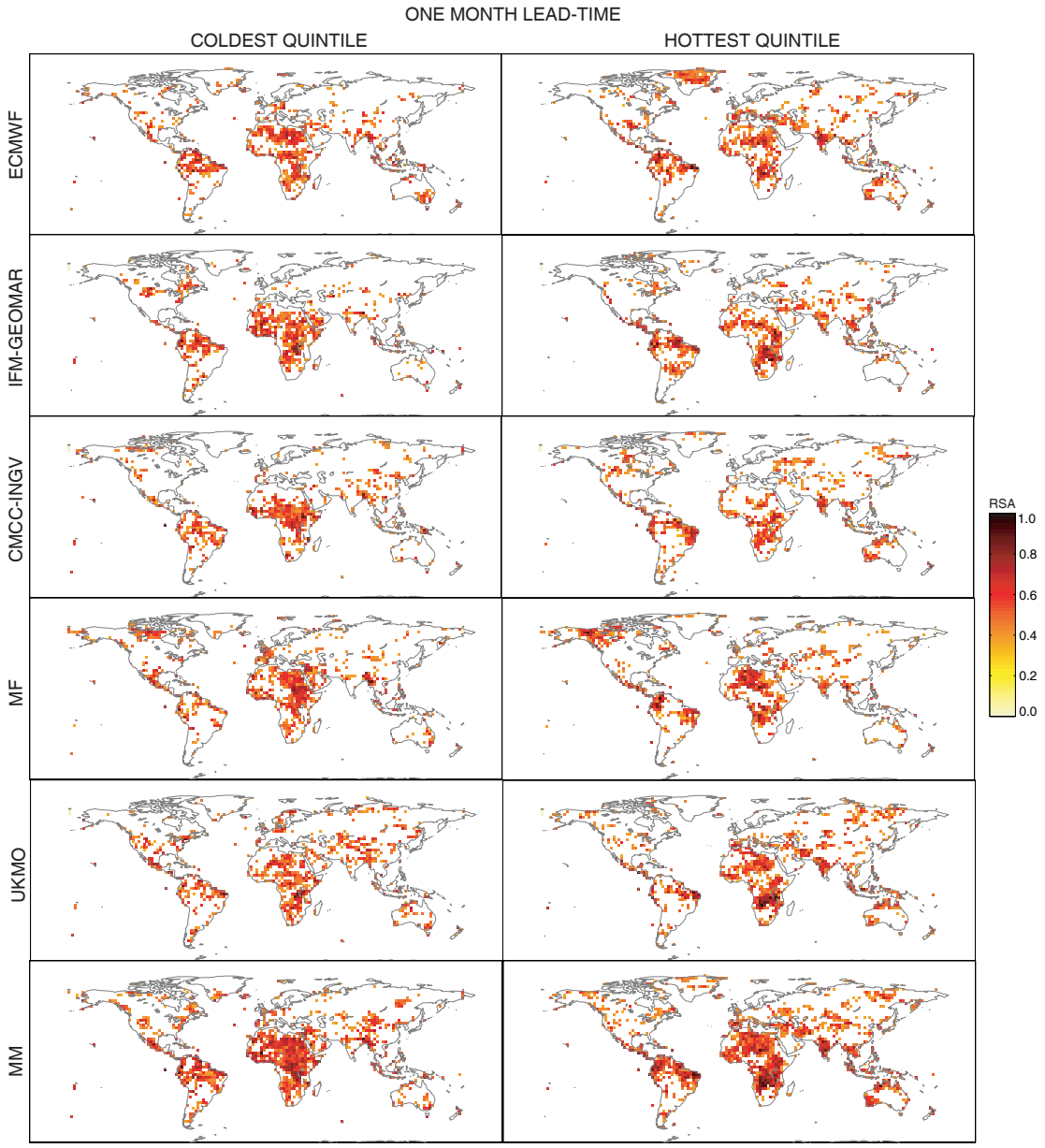

Figure 24. Left column: Models (rows) skill for the coldest quintile in JJA (period 1961-2000) at one month lead-time (trended data). Right column: Idem but for the hottest quintile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.

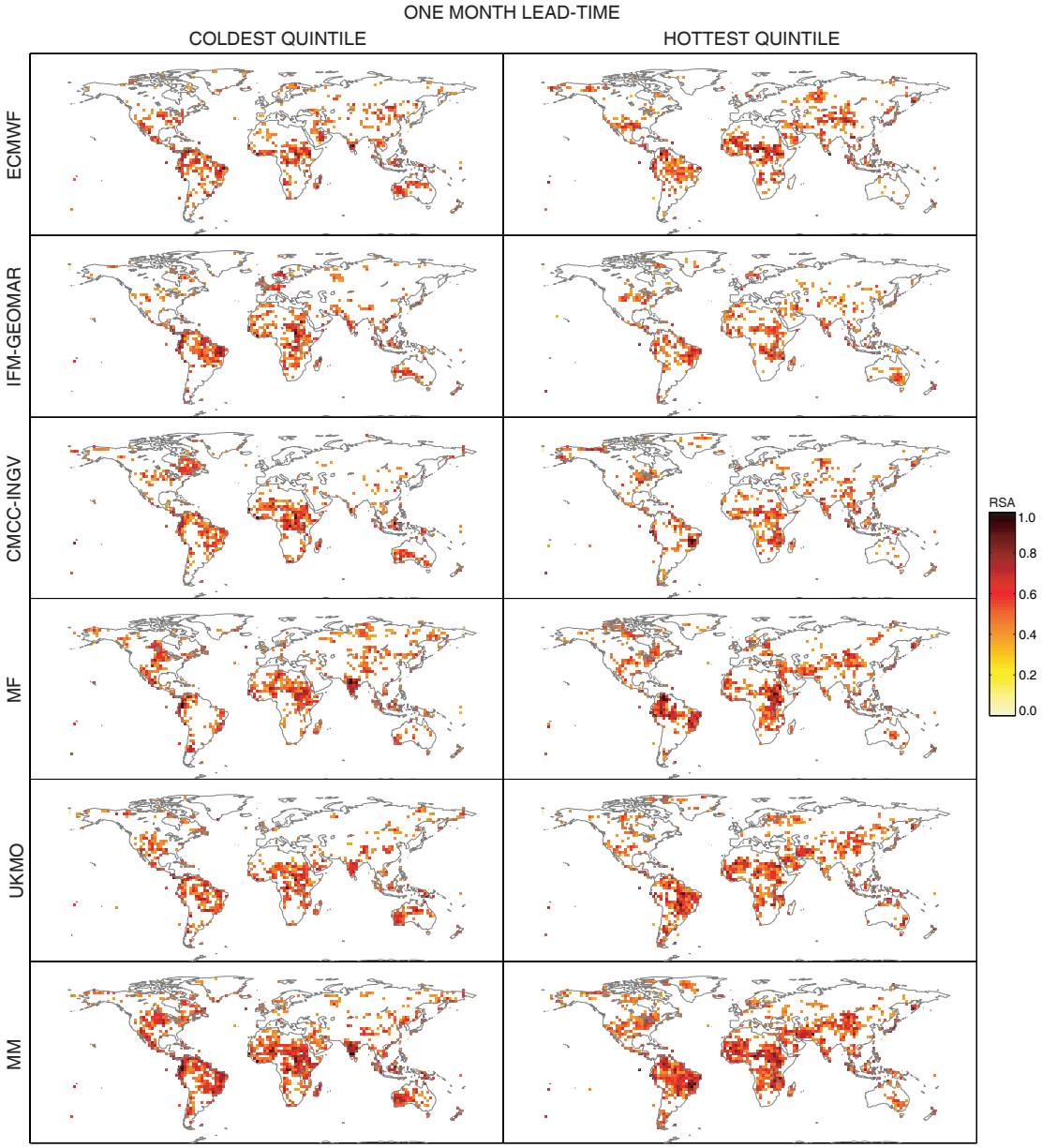

Figure 25. Left column: Models (rows) skill for the coldest quintile in SON (period 1961-2000) at one month lead-time (trended data). Right column: Idem but for the hottest quintile. Only points statistically significant at a confidence level of 90% (bootstrapping with 1000 samples) are shown.