

VERIFICATION OF SIX OPERATIONAL SEASONAL FORECAST SYSTEMS OVER EUROPE AND NORTHERN AFRICA

Eroteida SÁNCHEZ GARCÍA, José VOCES ABOY, Ernesto RODRÍGUEZ CAMINO
Agencia Estatal de Meteorología
esanchezg@aemet.es, jvocesa@aemet.es, erodriguezc@aemet.es

ABSTRACT

Verification scores of temperature and precipitation seasonal forecasts from six different operational systems for different seasons, variables and sub-regions over Europe and Northern Africa are computed based on available hindcasts. The operational systems here compared correspond to the most recent versions of seasonal forecast systems from the following Global Producing Centres: i) European Centre for Medium-Range Weather Forecasts (ECMWF), ii) Météo-France (MF), iii) UK Met Office (UKMO), iv) USA National Center for Environmental Prediction (NCEP), v) Environment Canada (EC) and vi) Japan Meteorological Agency (JMA). Two different verification periods were considered: firstly, 1997-2009, the maximum common period for all models, and, secondly, the maximum available hindcast period for each model. Results show noticeable dependency on the selected verification period. As expected, scores for temperature are better than for precipitation, with large differences among regions. Windows of opportunity associated to certain seasons/variables/models/regions are clearly identifiable. The dependency of such windows on the verification period weaken conclusions and their applicability.

Keywords: Seasonal Forecasts Verification Europe Africa

CONTENTS

1. INTRODUCTION.....	3
2. DATA.....	4
3. METHODOLOGY.....	5
4. RESULTS.....	9
4.1. Verification period: 1997-2009	19
4.2. Verification period: maximum available period for each model	47
5. SUMMARY AND CONCLUSIONS.....	75
6. REFERENCES.....	76
ANNEX I: RESULTS FOR TEMPERATURE WHEN TREND IS REMOVED....	78

1. INTRODUCTION

The chaotic features of the atmosphere limit the predictability of deterministic weather forecasts up to 10-15 days. Beyond this range, the predictability of atmospheric conditions has only sense from a statistical point of view and therefore forecasts must be expressed in probabilistic terms (Murphy and Winkler 1984). The main sources of uncertainty of forecasts at seasonal time scales come from the insufficient knowledge of initial conditions for the climate system and the lack of accuracy of climate models (Curry and Webster 2011, Knutti 2010, Slingo and Palmer 2011). The first source of uncertainty is explored using ensemble techniques based on independent forecasts from slightly different initial conditions (Gneiting and Raftery 2005; Palmer 2000). The second source of uncertainty is estimated, among other techniques, by combining different climate model integrations (Weisheimer et al. 2009). Predictability at seasonal time scale is highly dependent on particular atmospheric and oceanic modes of variability, regions, seasons and variables. Operational seasonal forecasts are frequently circumscribed to temperature and precipitation (Kirtman and Pirani 2008). The weak atmospheric predictability in mid-latitudes, and in particular over the Mediterranean region, has given preference to simple and robust seasonal forecasts based on terciles (Doblas-Reyes 2010).

This primary purpose of this report is to gain knowledge about the skill of the here considered models as a function of the season, variable and region in order to improve the operational seasonal forecast activities in different regions over Europe and Northern Africa, in particular, for the operational production of the Mediterranean Climate Outlook Forum (MedCOF). The seasonal forecast skill is evaluated for each region using two different periods of verification: firstly using the maximum common period for all models (1997-2009) and secondly using the maximum available period for each model, in order to identify windows of opportunity or circumstances with higher skill. These windows of opportunity may be linked to certain teleconnections, seasons, variables or specific forecast systems. The windows of opportunity can be originated by signals from several processes interacting constructively, but in many cases their reasons for such occurrence are still unclear. This report is an update of the previous report “*Calibration and combination of seasonal forecast over Southern Europe*” (Sánchez, E. 2014, AEMET Tech. Report, NIPO: 281-14-014-X) that included a selection of verification scores for different domains, seasons and variables, with additional models and verification domains. In particular, direct outputs from the following six models: *European Centre for Medium-Range Weather Forecasts (ECMWF) system 4*, *Météo-France system 5*, *UK Met Office system 9 (GloSea5)*, *National Center for Environmental Prediction (NCEP) system version 2*, *Canadian Seasonal to Inter-annual Prediction System (CanSIPS)* and *Japanese Seasonal Forecasting System 2* have been verified for fourteen domains defined over Europe and Northern Africa.

This technical note is organized as follows: Section 2 is a comprehensive description of data sources from both seasonal forecast models and observations while Sections 3 summarizes methodology and verification scores here applied. Finally, conclusions and way forward are discussed in Section 5.

2. DATA

The Full Data Reanalysis (V7) from the Global Precipitation Climatology Centre (GPCC) operated by the Deutsche Wetterdienst (DWD) from 1901 to 2013 has been used for the precipitation observational data and the ERA-Interim Reanalysis from the European Centre for Medium-Range Weather Forecast (ECMWF) from 1979 up to present for temperature observational data. Both of them have been interpolated to a common $2.5^\circ \times 2.5^\circ$ lat/lon grid.

A set of retrospective forecasts (or hindcasts) provided by the following centres have been used for their verification at seasonal time scales:

- The ECMWF *system 4* (S4) consists of the atmosphere Integrated Forecast System (IFS, version Cy36r4) at TL255 resolution (80 km grid point resolution) coupled with the ORCA1 configuration of the Nucleus for European Modelling of the Ocean (NEMO). The IFS has 91 levels and includes the whole stratosphere. Ocean initial conditions come from an assimilation system based on an advanced multivariate variational analysis with bias adjustments. Atmosphere and land surface initial conditions come from a mixture of ERA Interim (Dee et al. 2011) and ECMWF operations, and an offline run of the HTESSEL surface model (Kim et al. 2012, Molteni et al. 2011).
- The Météo-France system 5 (MF5) consists of the ARPEGE-Climat version 4 (Action de Recherche Petite Echelle Grande Echelle) for the atmospheric component coupled with ORCA, developed by LOCEAN, for the ocean model. The ocean initial conditions are prepared by MERCATOR. The atmospheric model has a horizontal resolution of 0.75° (TL255 truncation) and 91 levels allowing an explicit representation of the stratosphere along with ozone and non-orographic gravity wave drag.
- The UK Met Office system 9 (GLOSEA5) has NEMO as ocean model with a spatial resolution of $0.25^\circ \times 0.25^\circ$ with 75 levels, and the atmosphere model is GEM3 GA3.0. The operational setup N216L85 corresponds to approximately 50 km horizontal resolution and 85 levels
- The National Center for Environmental Prediction system version 2 (CFSv2) has an atmospheric component with a spatial resolution of 100 km and 64 vertical levels (Kim et al. 2012, Saha et al. 2013, Yuan et al. 2011). The ocean component is the Geophysical Fluid Dynamics Laboratory Modular Ocean Model (MOM4) version 4 with horizontal resolution of 0.5° , refined at 0.25° between 10°N and 10°S , and 40 vertical levels.
- The Japanese Meteorological Agency seasonal forecast system (JMA System 2) comes from JMA/MRI-CPS2. Atmospheric resolution is TL159 (approx 110km grid spacing), with 60 levels and a model top at 0.1 hPa. Atmosphere and land initial conditions come from JRA-55 and ocean initial conditions come from MOVE/MRI.COM-G2.
- The Canadian Seasonal to Interannual Prediction System (CanSIPS) from the Canadian Meteorological centre (Montreal, CMC) has CanAM4 as atmospheric model

with a resolution of T63/L35 ($\approx 2.8^\circ$ spectral grid), and CanOM4 as ocean model with a resolution of $1.41^\circ \times 0.94^\circ \times L40$.

Brief descriptions of technical specifications for the above mentioned forecast systems are available in the Lead Centre for Long-Range Forecasts repository (<https://www.wmolc.org>).

3. METHODOLOGY

Seasonal forecasts of temperature and precipitation from the six seasonal systems here considered are verified using both deterministic and probabilistic skill scores. Statistical significance of all computed scores has been quantified by the p-value estimated using a bootstrapping non-parametric method (Wilks 2006).

The only deterministic skill score computed for both temperature and precipitation is the correlation between the predicted and the observed mean value of anomalies over the different land domains. The score was computed for 12 different consecutive three-month periods and always for 1 month lead time 1. By way of clarification, lead time refers to the period of time between the issue time of the forecast and the beginning of the forecast validity period. Long-range forecasts based on all data up to the beginning of the forecast validity period are said to be of lead zero. The period of time between the issue time and the beginning of the validity period will categorize the lead. For example, a winter seasonal forecast issued at the end of the preceding summer season is said to be of one season lead. A seasonal forecast issued one month before the beginning of the validity period is said to be of one month lead.

The following probabilistic skill scores have been also computed for the same variables (temperature and precipitation), for the same 12 different three-month periods and for 1 month lead time: Ranked Probability Skill Score (RPSS) for terciles, and Relative Operating Characteristic (ROC) area and Brier Skill Score (BSS) for two events (upper/lower tercile). A complete definition of these scores can be found in Wilks (2006).

By way of summary, the Ranked Probability Skill Score (RPSS) is a generalization of Ranked Probability Score (RPS) based on a reference forecasting system. The RPS averages squared “error” in the cumulative probabilistic forecasts. Positive values of RPSS indicate more skill than the reference system, usually the climatology.

ROC curves measure discrimination and skill. If the category of interest is above-normal, the score based on the ROC area indicates the probability of successfully discriminating above-normal observations from normal and below-normal observations. The ROC area ranges from 0% to 100%, with a score of 50% representing no skill, 100% indicating perfect discrimination, and 0% indicating perfectly bad discrimination. It is important to stress that ROC curves are measuring only the discrimination ability between two possible results, but it is not informative about reliability since it is not sensitive to bias.

The Brier Score (BS) is the most common verification method for probabilistic forecasts, as it, has a mathematical structure similar to the Mean Square Error (MSE). BS measures the difference between the forecast probability of an event (p) and its occurrence (o), expressed as 0 or 1, depending on whether the event has occurred or not. As with RMSE, the BS is negatively orientated, i.e. the lower, the “better”. The Brier Skill Score (BSS) is

conventionally defined as the relative probability score compared with the probability score of a reference forecast.

All scores have been computed for two different periods of verification: i) the maximum common period for all models (1997-2009), and ii) the maximum available period for each model. Hindcast periods available are also depending on variables (See Table 1 and Table 2)

TEMPERATURE

MODEL	PERIOD	Number of years
Can	1982 – 2010	29
CFSv2	1983 – 2015	33
GloSea5	1997 – 2009	13
JMA2	1982 – 2015	34
MF5	1992 – 2014	23
S4	1982 – 2015	34

Table 1. - Maximum available verification periods for temperature

PRECIPITATION

MODEL	PERIOD	Number of years
Can	1982 – 2010	29
CFSv2	1983 – 2012	30
GloSea5	1997 – 2009	13
JMA2	1982 – 2012	31
MF5	1992 – 2012	21
S4	1982 – 2012	31

Table 2. - Maximum available verification periods for precipitation

The same collection of verification scores have been computed for fourteen areas covering Europe and Northern Africa (see Table 3 and Fig.1).

Area	NW Corner	SE Corner
ALPINE	48N/5E	43N/16E
ALGERIA/TUNISIA	37.5N/4W	19N/12E
BALKANS	48.5N/13E	34.5N/29E
CENTRAL-EUROPE	55N/2.5E	46N/19E
EAST-EUROPE	55N/14E	44N/40E
EAST MED	38N/21E	29N/39E
FRANCE	52N/6.4W	41N/10E
GREAT BRITAIN	59N/11W	50N/2E
IBERIA	44N/10W	36N/4.5E
ITALY	47N/6.5E	36N/18.5E
LIBYA/EGYPT	33N/9E	20N/36E
MOROCCO	36N/17W	21N/1W
SCANDINAVIA	66N/5E	54N/32E
TURKEY	42N/26E	36N/50E

Table 3 – Coordinates (latitude, longitude) of the selected domains

All scores have been calculated using three monthly anomalies values. From the observational monthly data, three monthly anomaly values of precipitation and temperature at $2.5^\circ \times 2.5^\circ$ horizontal resolution were computed to verify seasonal models outputs.

The anomalies of the different prediction systems, computed as the difference between the forecasted and climatological values for each system, using the corresponding verification period as climatology period, are obtained by cross-validated forecasts on data not used in the estimation, i.e., the year to be forecast is removed from the data set.

All available hindcast ensemble members corresponding to each verification period have been used in the computation of lower and upper terciles. Terciles for observational data were also computed over the same corresponding period. All scores have been also calculated removing the temperature trend, in order to know how this fact may affect forecasts skill (See Annex I). Finally, as data were provided in latitude-longitude geographical coordinates, grid point values are weighted by the cosine of latitudes.

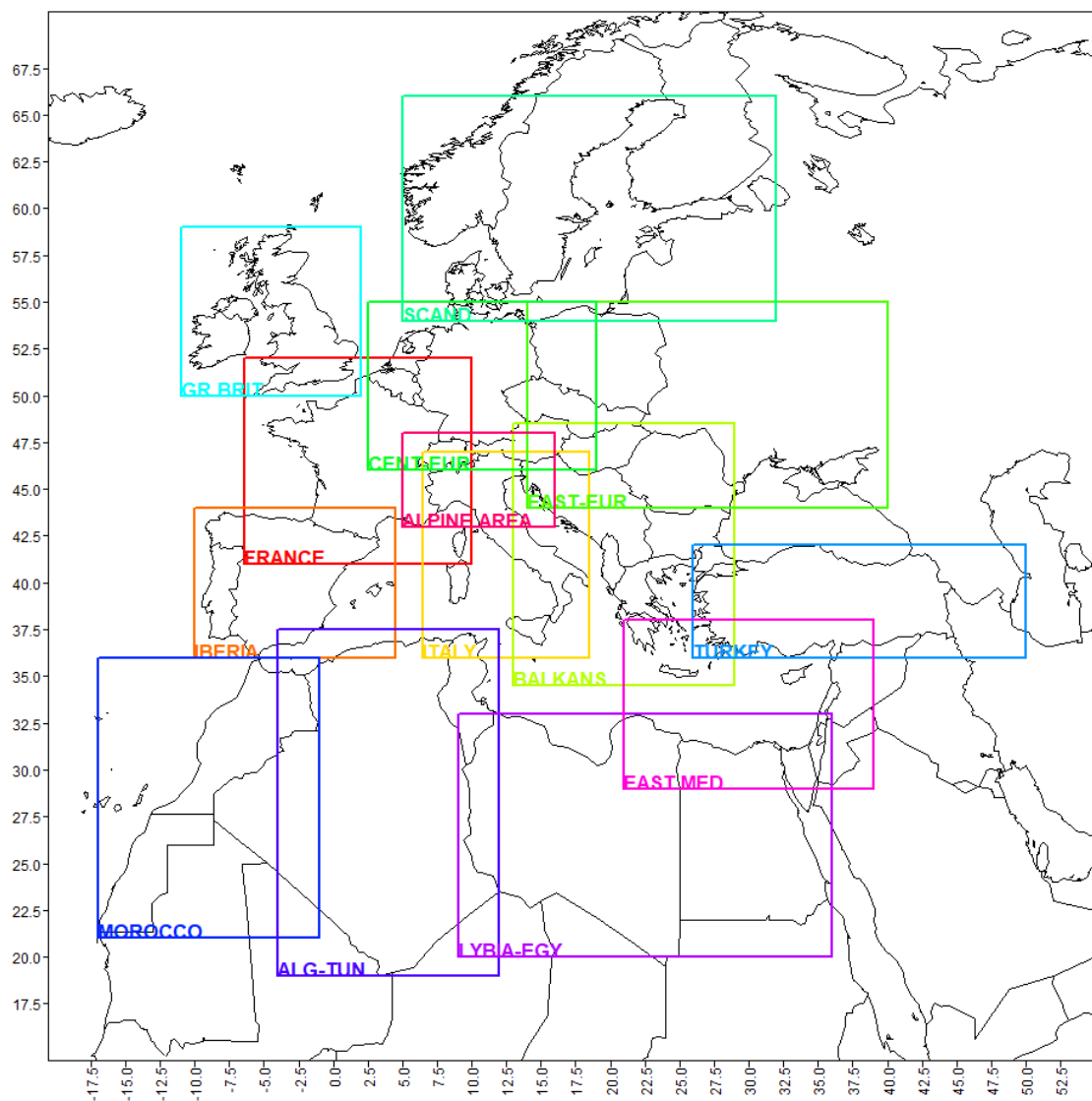


Fig. 1 - Selected land domains over Europe and Northern Africa

4. RESULTS

The four skill scores described in the previous section have been calculated taking into account only land grid points on each selected domain - sea grid points were removed for verification purpose-. Calculated values for each score and over each selected domain are displayed using tables referred to anomalies of temperature and precipitation for 12 different three-month periods and for lead-time 1 (see Tables 4 to 59).

Results removing trend for temperature are additionally shown in Annex I (Tables 60 to 87). Results are individually discussed for each of the selected domains following scores summarized in Tables 4 to 59.

Alpine Area:

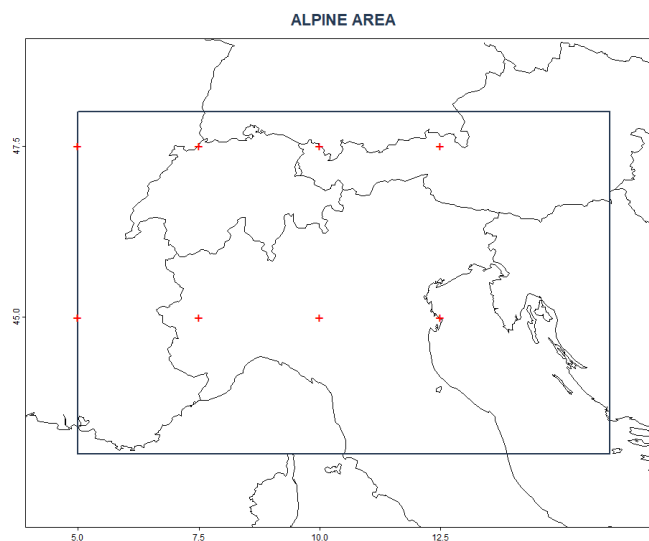


Fig. 2- Grid points over the ALPINE AREA domain

Temperature shows some significant values for all calculated scores and for some specific 3 months period or forecast system using the common verification period 1997-2009. When the maximum available period for each model was applied for the computation of verification scores, most systems tend to show an increase of the skill during spring-summer (BSS and RPSS greater than 0, ROC area values greater than 0.6, or even 0.7 in some cases). However, if temperature trend is removed before calculating verification scores, values obtained as expected are remarkably smaller for models having longer hindcasts (S4, JMA2, Can, CFSv2).

Scores for precipitation are generally much worse. Correlation coefficients, BSS and RPSS are very small ($r < 0.3$ for much of months/models and BSS/RPSS < 0 except for few cases). Only (lower and upper) ROC area scores are better than climatology (greater than 0.5) for spring and autumn (reaching almost 50% of total cases when maximum available period for each model is used).

Algeria-Tunisia

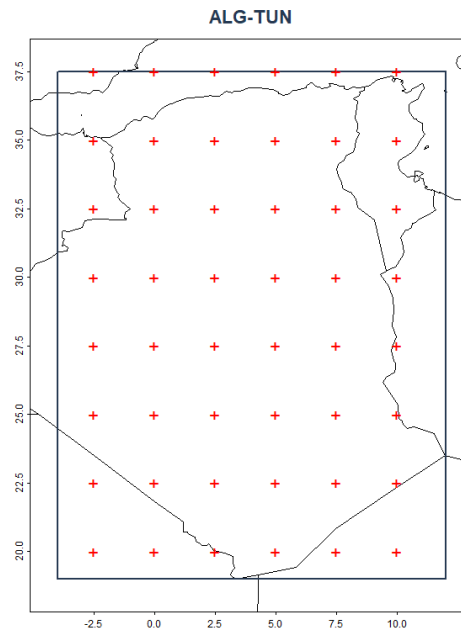


Fig. 3.- Grid points over the ALGERIA/TUNISIA domain

Scores for temperature show a remarkable skill for most months (overall in spring, summer and autumn) when the maximum available hindcast verification period (CSFv2, Can, JMA2 and S4 models) is used. When temperature trend is not removed, correlation coefficients are greater than 0.5 in some cases, BSS greater than 0.2 and ROC area bigger than 0.7. Again when trend is removed, values decrease. However, in about 75% of the figures appearing in the tables, discrimination is better than climatology, being the best values in summer.

With respect to precipitation, tables show less reliability than for temperature, however discrimination is better than climatology in about 75% of cases (spring-summer-autumn).

Balkans:

Scores for temperature show good significant values for spring-summer seasons (some months or models even show correlation coefficients greater than 0.5, BSS greater than 0.2 and ROC values above 0.7). As for the rest of domains, when using the common verification period 1997-2009 or removal of temperature trend, scores decrease for some models/months.

For precipitation, correlation coefficient, BSS and RPSS scores are generally not so good than for temperature showing low values over most months/models (even a few cases with $r < 0.2$ and BSS and RPSS < 0). Only ROC area (lower and upper) scores are better than climatology (greater than 0.5) in about 50% of total cases (most of them in spring and summer).

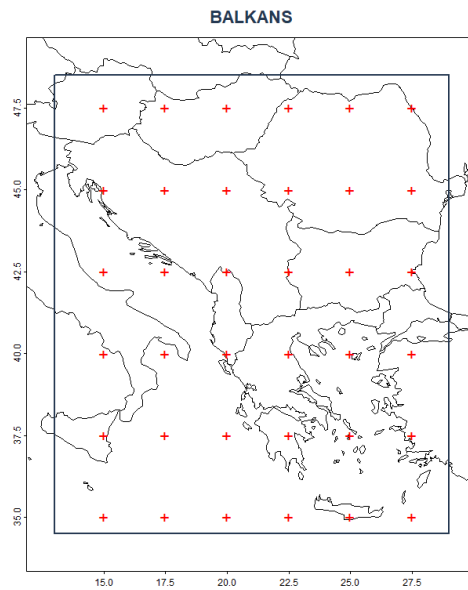


Fig. 4.- Grid points over the BALKANS domain

Centre Europe:

Temperature shows some significant good values for all calculated scores in spring-summer seasons (correlation coefficients greater than 0.5, BSS greater than 0.2 and ROC area values above 0.7 for some months/models), but not generalizable to all models. Again, skill decreases when removing temperature trend or using a short number of years to verify, as in the remaining domains.

For precipitation, skill is not noticeably better than climatology.

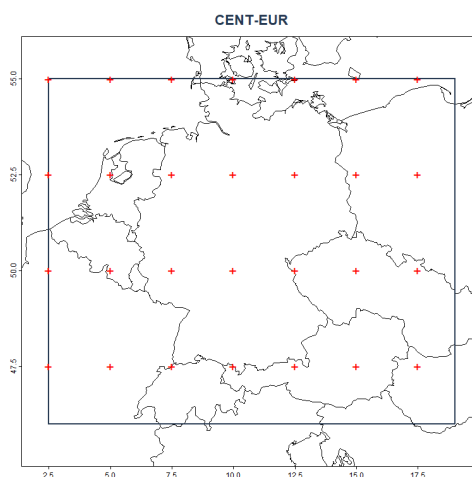


Fig. 5.- Grid points over the CENTRE EUROPE domain

East Europe:

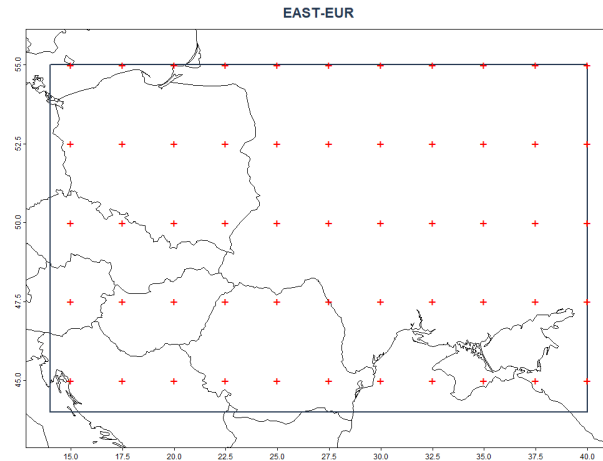


Fig. 6.- Grid points over the EAST EUROPE domain

As happened for Centre-Europe domain, scores are better in spring-summer seasons (even in some winter months). Scores for temperature in the spring (MAM) period are noticeably good in all cases (removing or not trend and using both verification periods) and for all models (correlation coefficient around 0.6-0.7, BSS greater than 0.3, ROC areas over 0.86).

East Mediterranean:

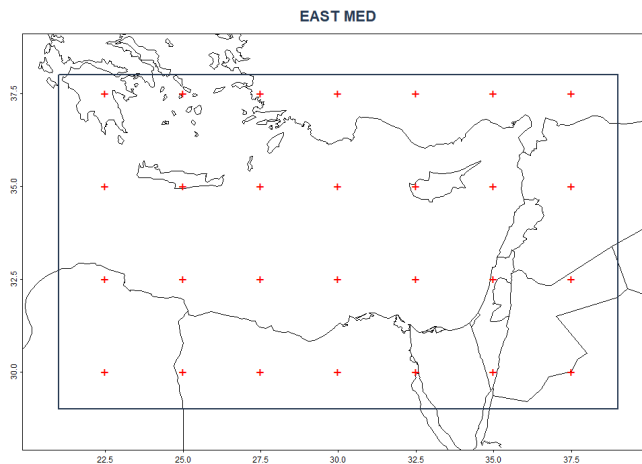


Fig. 7.- Grid points over the EAST MEDITERRANEAN domain

Scores for temperature show a remarkable skill using the largest verification period (CSFv2, Can, JMA2 and S4 models) in spring, summer, beginning of autumn and end of winter. Tables show correlation coefficient greater than 0.5 in some cases, BSS greater than 0.2 and ROC area larger than 0.7. These results are maintained even when removing trend, although scores somehow degrade, as happened in other domains.

With respect to precipitation, tables show remarkable scores for later summer and early autumn.

France:

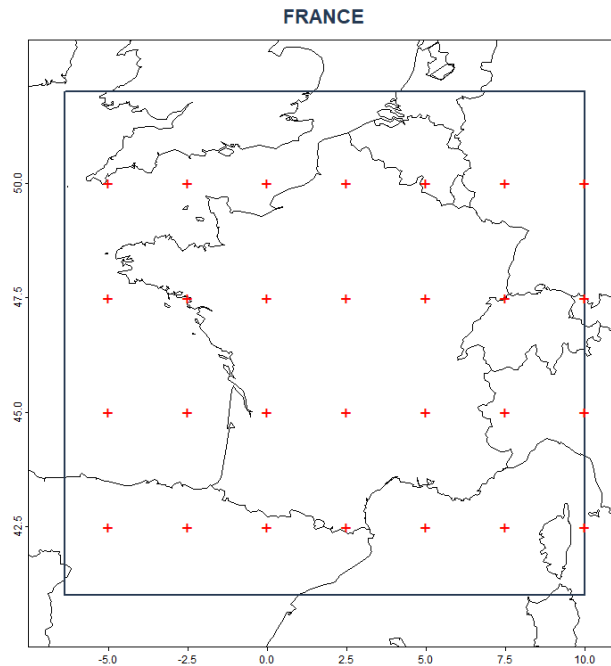


Fig. 8.- Grid points over the FRANCE domain

Temperature shows some significant values for all calculated scores and for some specific 3 months period or forecast system using the common verification period 1997-2009. When the maximum available period for each model was applied in the computation of verification scores, most systems tend to show an increase of the skill during spring-summer (BSS and RPSS greater than 0, ROC area values greater than 0.6, or even 0.7 in some cases). However, if temperature trend is removed before calculating verification scores, values obtained are remarkably smaller for those models having longer hindcasts (S4, JMA2, Can, and CFSv2).

Scores for precipitation are generally much worse. Correlation coefficients, BSS and RPSS are very small ($r < 0.3$ for most months/models and BSS/RPSS < 0 except for few cases). Only (lower and upper) ROC area scores are better than climatology (greater than 0.5) for spring and autumn (reaching almost 50% of total cases when maximum available period for each model is used).

Great Britain:

Tables for temperature show some significant values for all calculated scores and for some specific 3-months period or forecast system using the common verification period 1997-2009 (models have generally more skill in winter season). When using the maximum available hindcast period for each model, skill increases for most systems and 3-month periods (reaching in some cases $r > 0.5$, BSS > 0.2 and ROC area > 0.8). However, if temperature trend is removed, scores slightly degrade for some periods/models, but not as a general rule (even scores improve in some cases, for instance, correlation for winter months).

Scores for precipitation, as usual, are generally worse than for temperature. Correlation coefficients are higher than 0.3 in a few cases, whereas BSS and RPSS are very small (BSS

and $RPSS < 0$ except in a few cases). Again only ROC area (lower and upper) scores are better than climatology (greater than 0.5) for more than 50% of total cases. Skill in winter is slightly better than in other seasons.

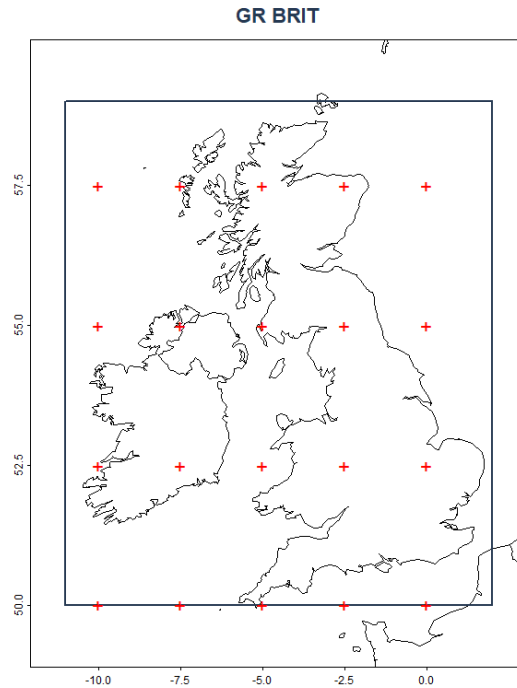


Fig. 9.- Grid points over GREAT BRITAIN domain

Iberia:

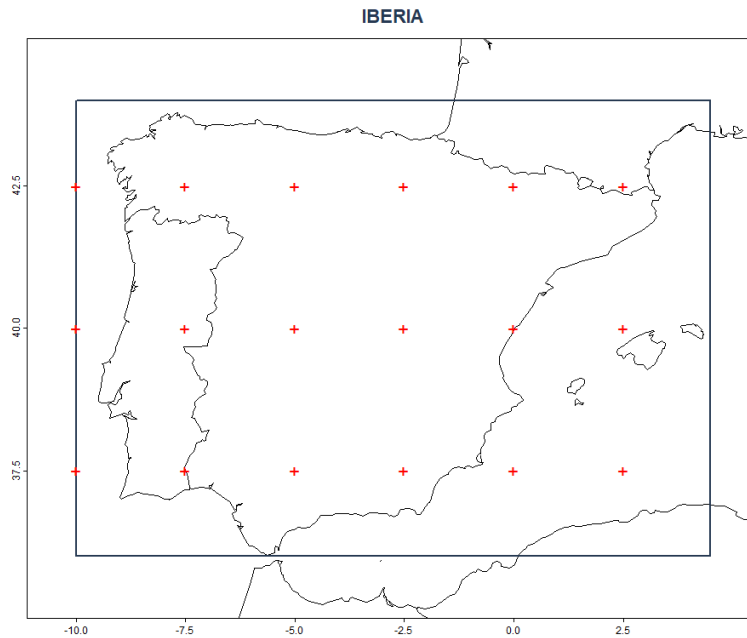


Fig. 10.- Grid points over IBERIA domain

As regards to temperature, all calculated scores show some significant values for some specific 3-month periods or forecast system using the common verification period 1997-2009.

If the maximum available verification period for each model is used, skill increases for most systems during the spring-summer (BSS and RPSS greater than 0, ROC area greater than 0.6, even 0.7 in some cases). However, if temperature trend is removed before calculating scores, a clear degradation is noticeable for models with longer hindcasts periods (S4, JMA2, Can, CFSv2) due to the clear warming trend during last decades.

Tables show for precipitation very low skill in terms of correlation coefficient, BSS and RPSS ($r < 0.3$, BSS/RPSS < 0 except for a few cases mostly in summer/autumn). Again ROC area (lower and upper) scores are better than climatology (higher than 0.5) in about 60-70% of cases depending on the verification period used.

Italy:

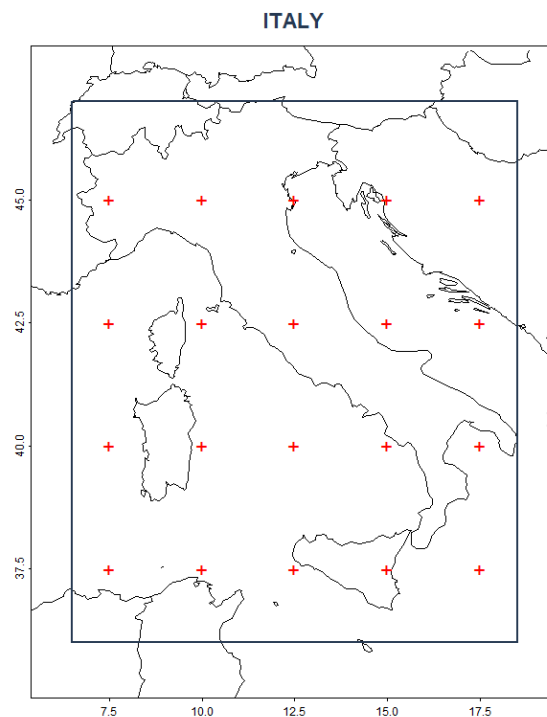


Fig. 11. - Grid points over the ITALY domain

Scores for temperature show good significant values for spring-summer seasons (some months or models even show correlation coefficients greater than 0.5, BSS greater than 0.2 and ROC values above 0.7). As for the rest of domains, when using the common verification period 1997-2009 or removal of temperature trend, scores decrease for some models/months.

For precipitation, correlation coefficient, BSS and RPSS scores are generally not so good than for temperature showing low values over most months/models (even a few cases with $r < 0.2$ and BSS and RPSS < 0). Only ROC area (lower and upper) scores are better than climatology (greater than 0.5) in about 50% of total cases.

LIBYA-Egypt:

Scores for temperature show a remarkable skill for most months (overall in spring, summer and autumn) when the maximum available hindcast verification period (CSFv2, Can, JMA2 and S4 models) is used. When temperature trend is not removed, correlation coefficients are greater than 0.5 in some cases, BSS greater than 0.2 and ROC area larger than 0.7. Again when trend is removed, values decrease. However, in about 75% of the figures appearing in the tables, discrimination is better than climatology; reaching the best values in summer.

With respect to precipitation, tables show less reliability than for temperature, however discrimination is better than climatology in about 75% of cases (spring-summer-autumn).

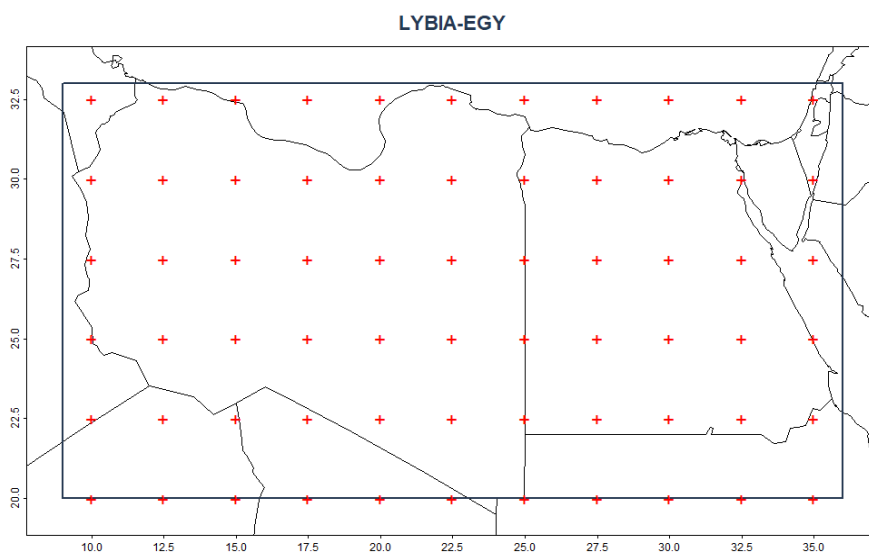


Fig. 12.- Grid points over LIBYA/EGYPT domain

Morocco:

With regard to temperature, scores are similar to those found in the ALGERIA domain, although values are not so high.

For precipitation, tables show poor reliability whereas discrimination is better than climatology in about 70% of total cases.

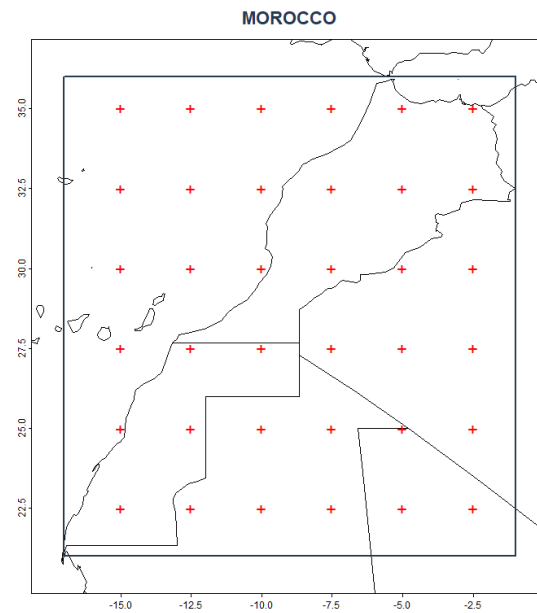


Fig. 13.- Grid points over MOROCCO domain

Scandinavia:

Both temperature and precipitation scores hardly differ from climatology over the Scandinavian domain. The only noticeable window of opportunity (temperature) appears in the MAM period, being this window consistent throughout scores, models and verification periods.

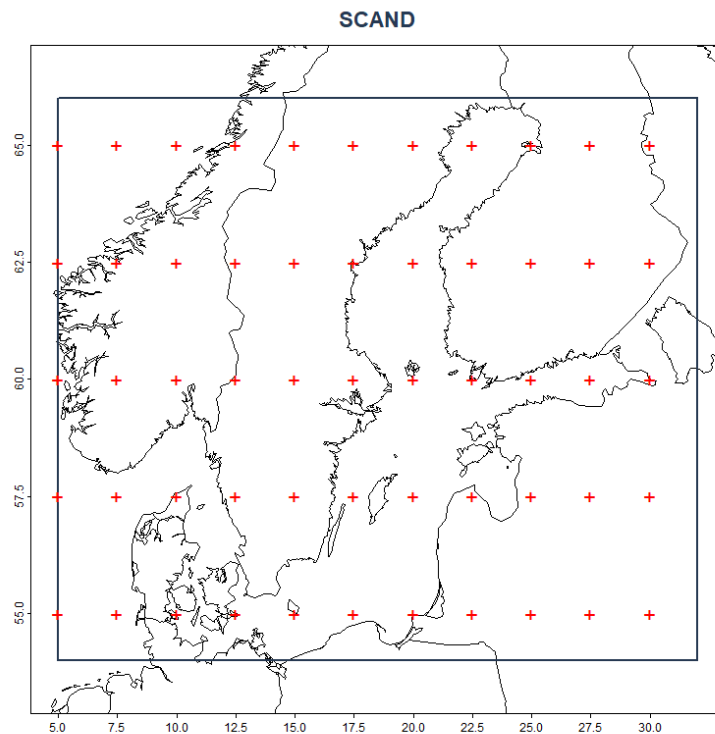


Fig. 14.- Grid points over SCANDINAVIA domain

Turkey:

Scores for temperature show remarkable skill (reaching r 0.73, RPSS 0.34, BSS 0.25 and ROC area 0.81). A clear wide window of opportunity appears for most models (covering end of winter, spring and summer). Removal of temperature trend generally causes some degradation, as in other domains, although the window is still noticeable.

Scores for precipitation are in general worse than for temperature. Correlation coefficient is higher than 0.3 in only few cases, whereas BSS and RPSS are generally very low (BSS and RPSS < 0 except for a few cases). However, ROC area (lower and upper) scores are better than climatology (higher than 0.5) in about 80% of figures appearing in tables.

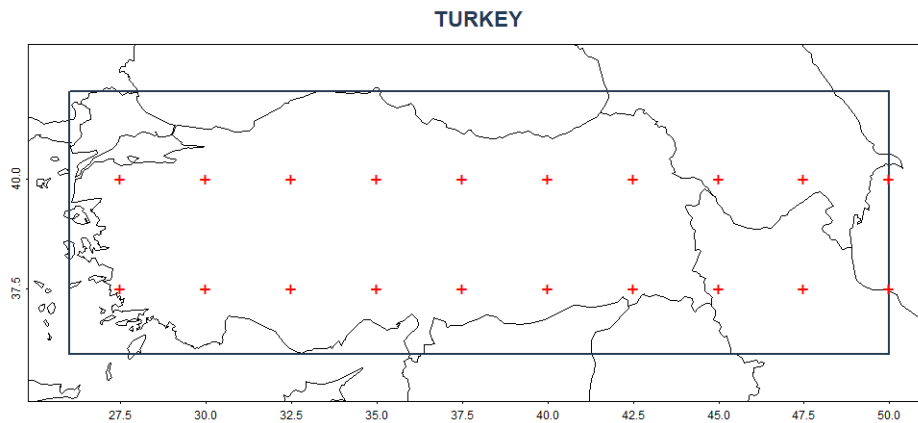


Fig. 15.- Grid points over TURKEY domain

4.1. Verification period: 1997-2009

4.1.a. Temperature

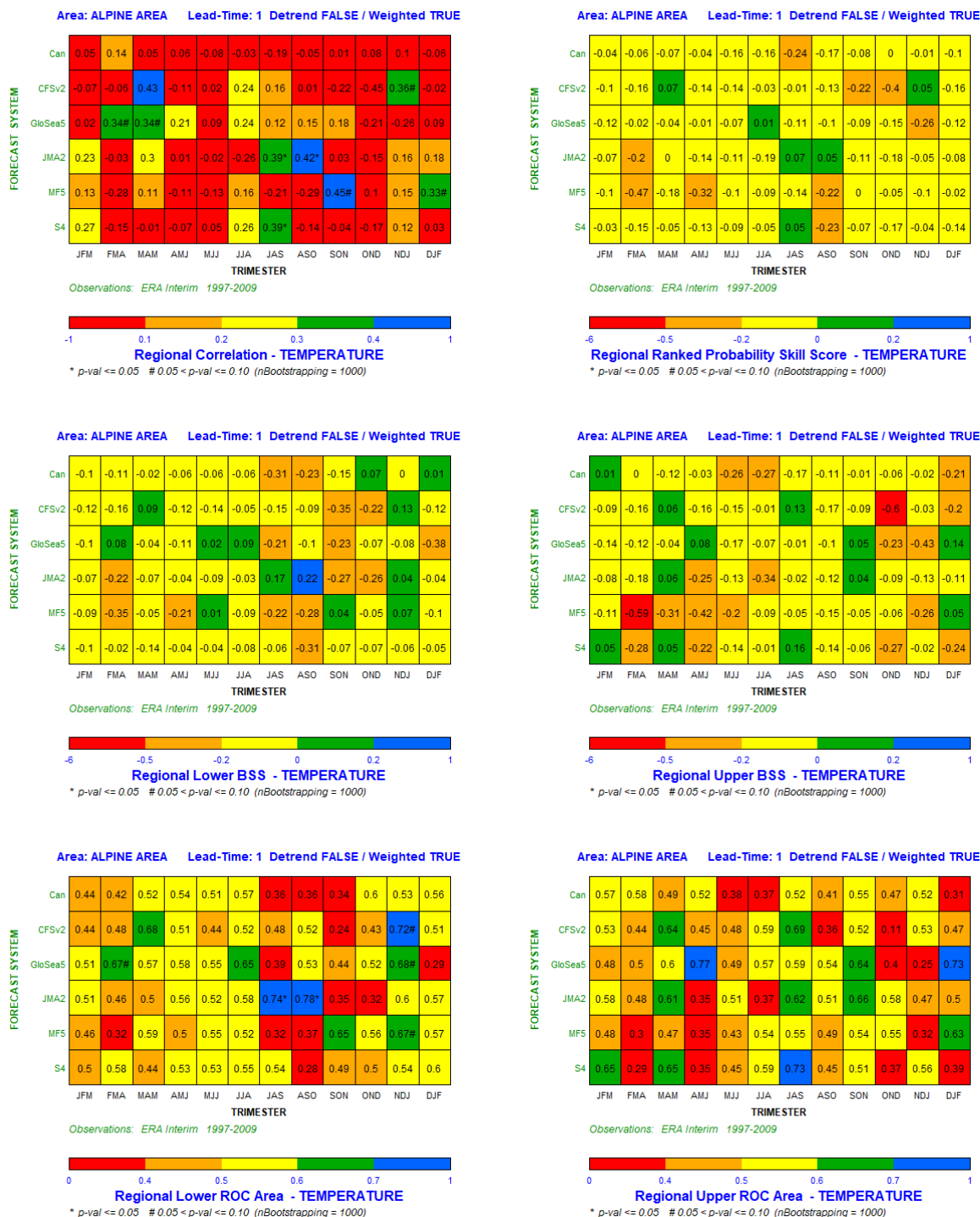


Table 4. Regional correlation coefficient, Ranked Probability Skill Score, Lower and Upper ROC Area and Brier Skill Score computed for temperature anomalies (without removing trend), for 12 different three-month periods and lead-time 1 over ALPINE AREA domain (verification period: 1997-2009). Three-month periods are shown in the X-axis and systems (S4, MF5, JMA2, GLOSEA5, CFSv2 and Can) are represented in the Y-axis (see text for description).

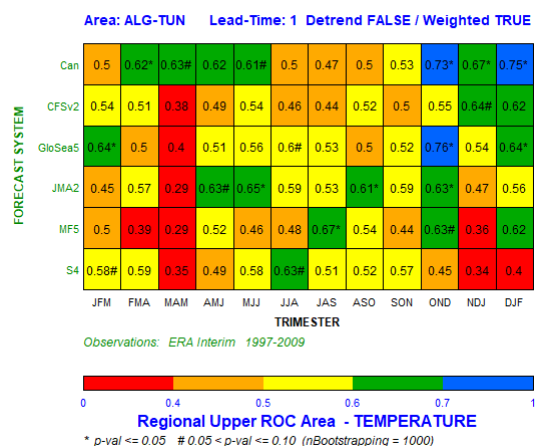
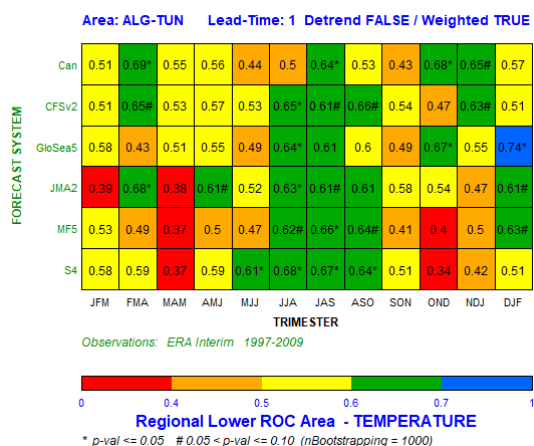
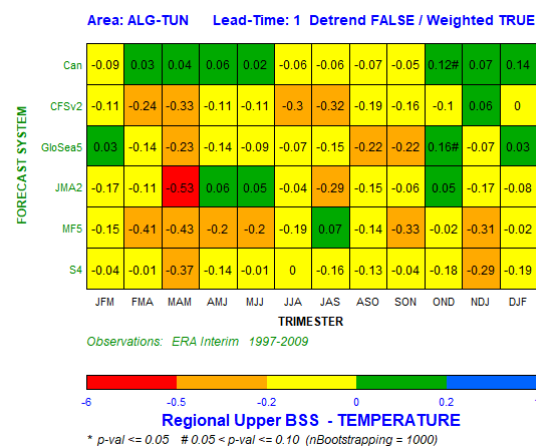
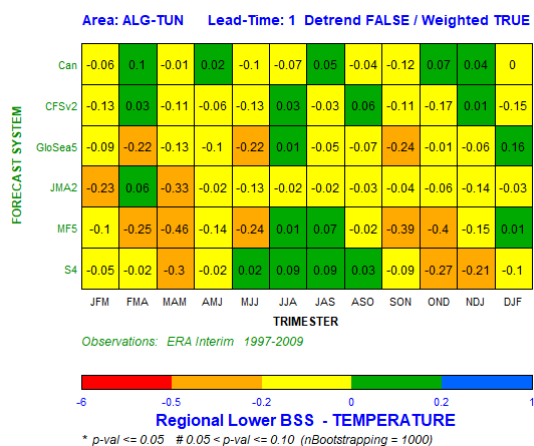
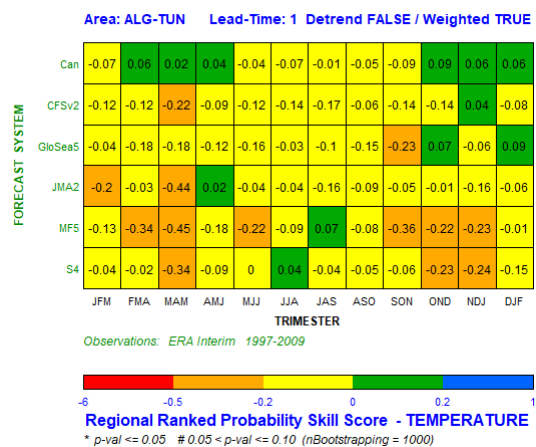
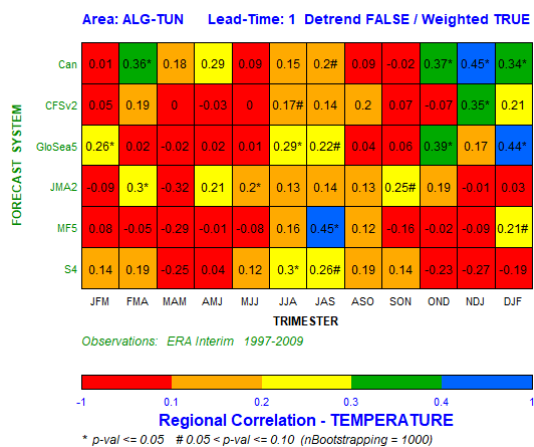


Table 5. The same as Table 4, but for ALGERIA/TUNISIA domain

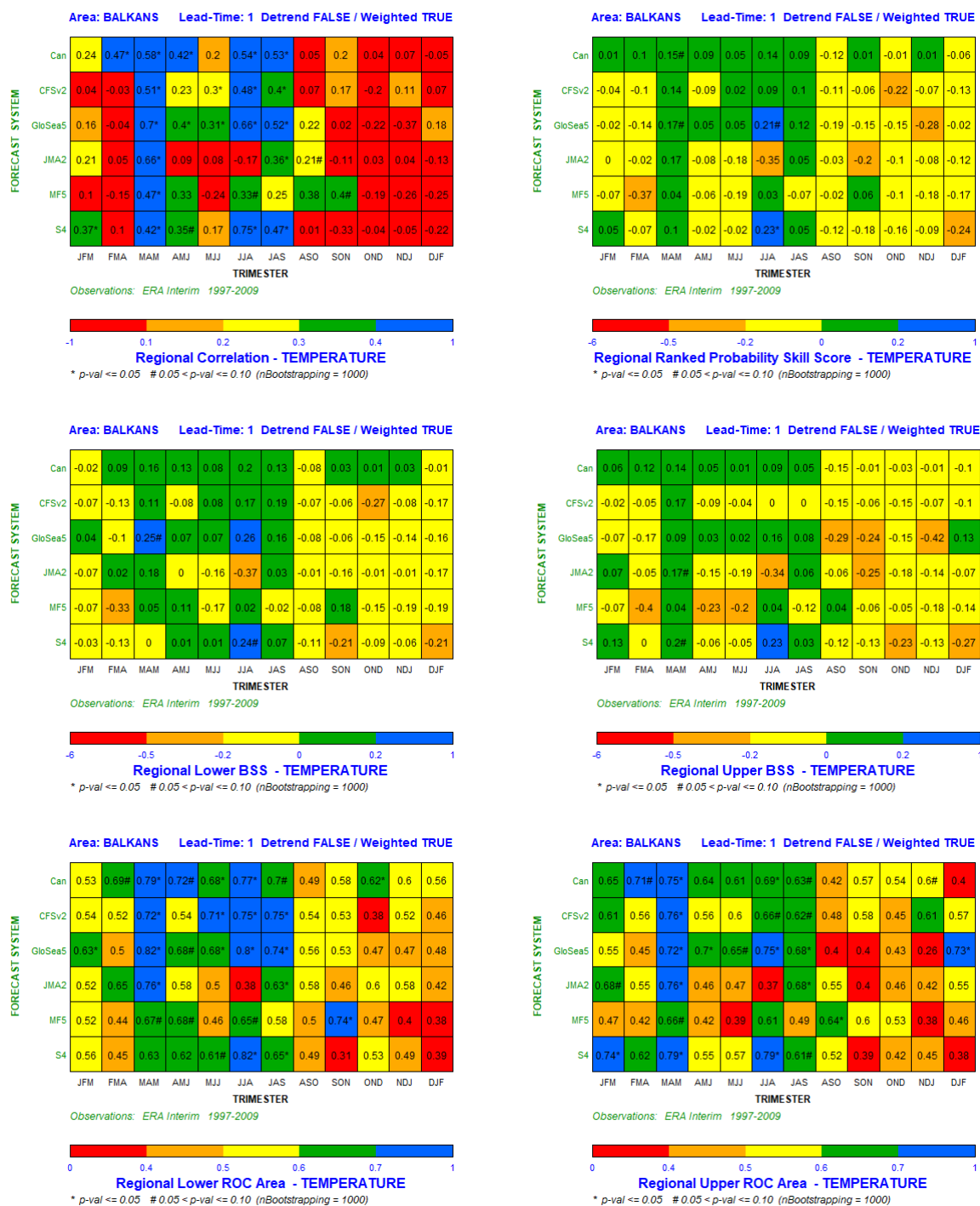


Table 6. The same as Table 4, but for BALKANS domain

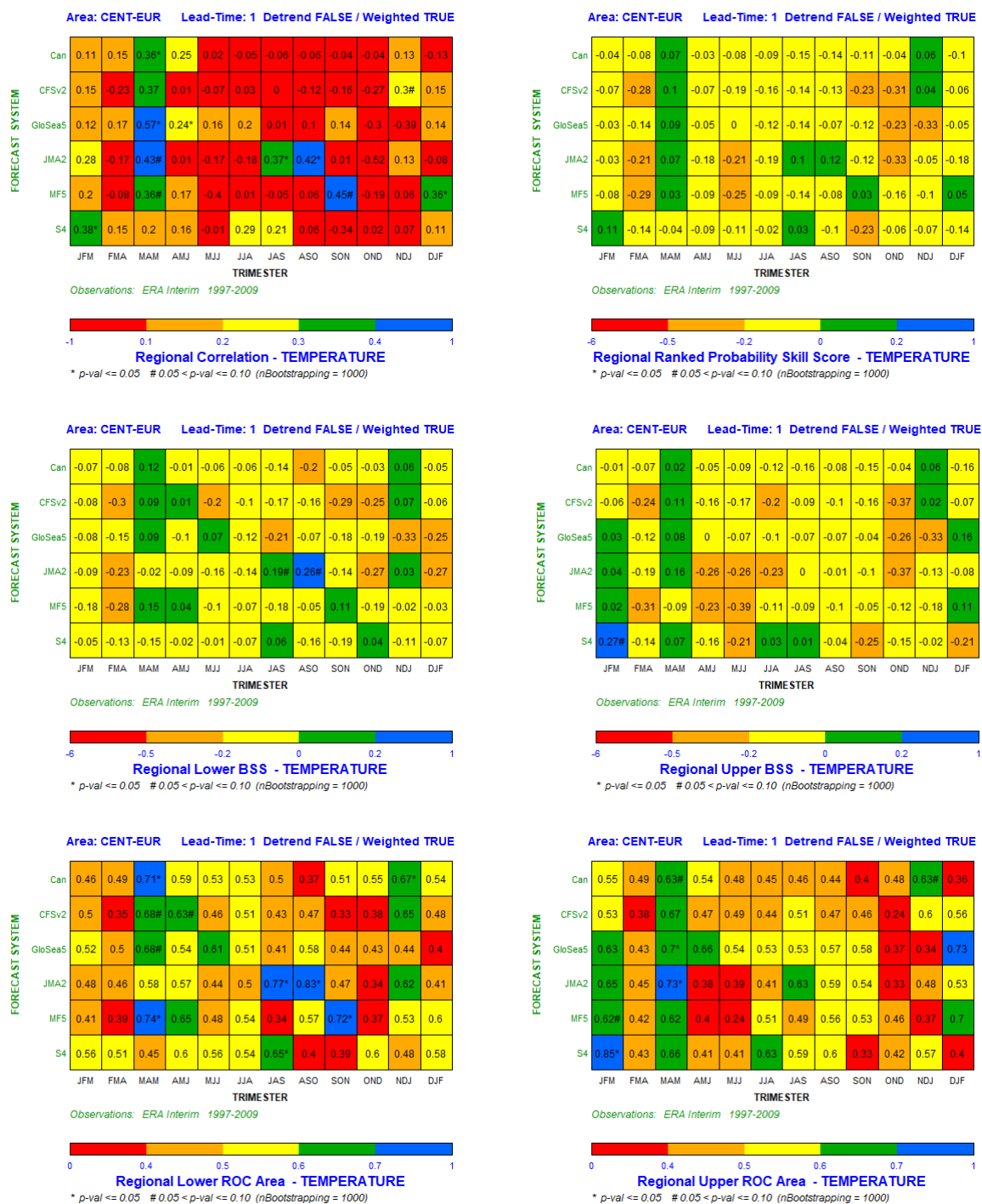


Table 7. The same as Table 4, but for CENTRE EUROPE domain

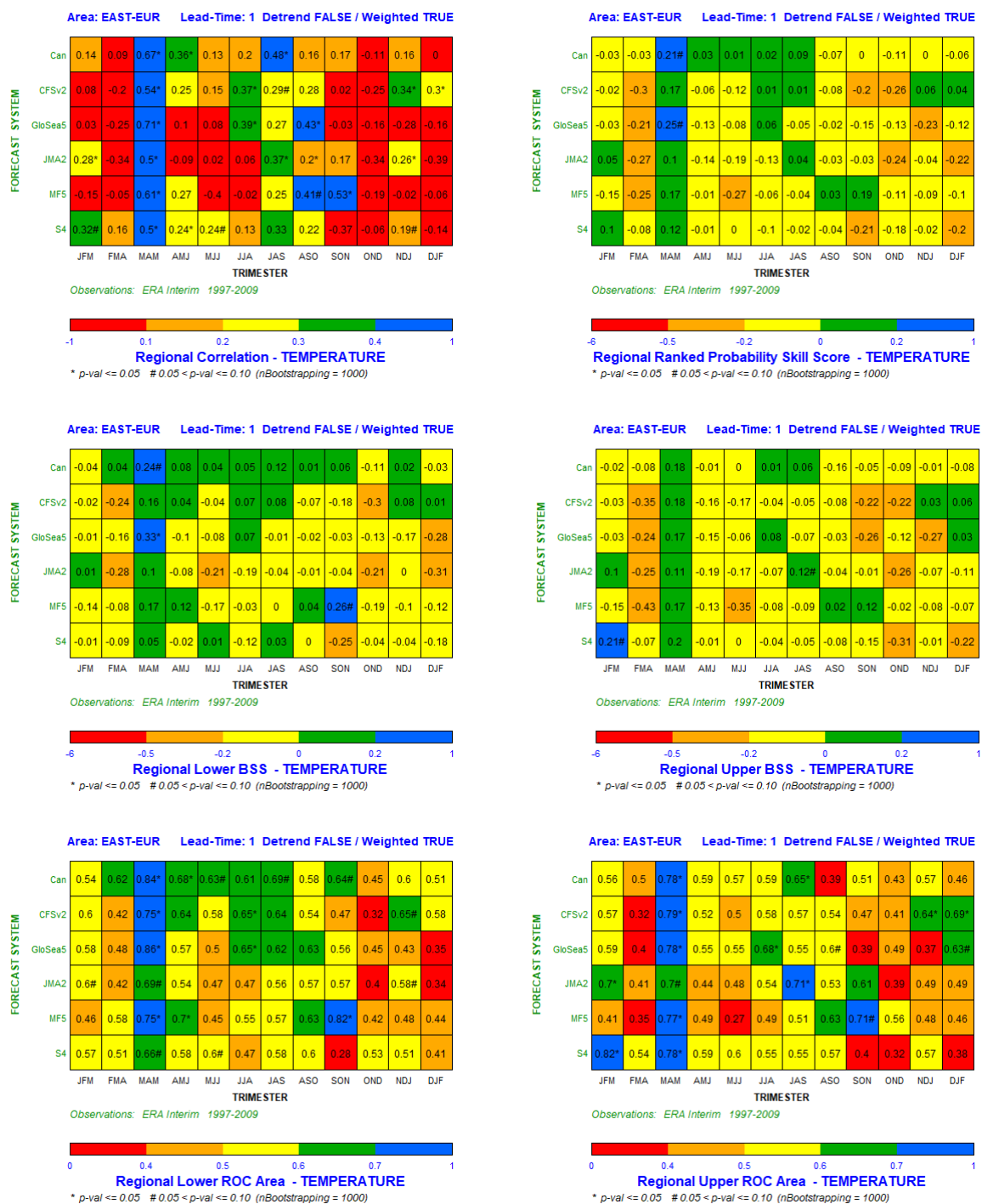


Table 8. The same as Table 4, but for EAST EUROPE domain

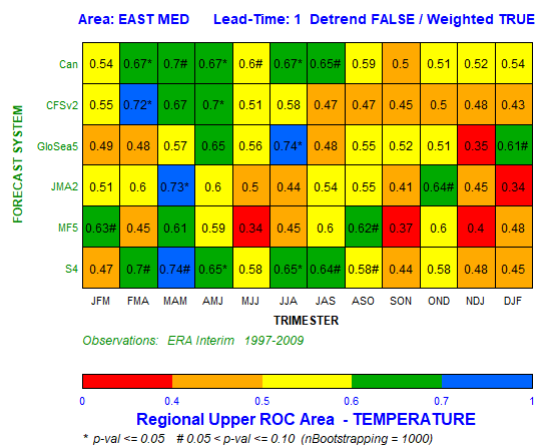
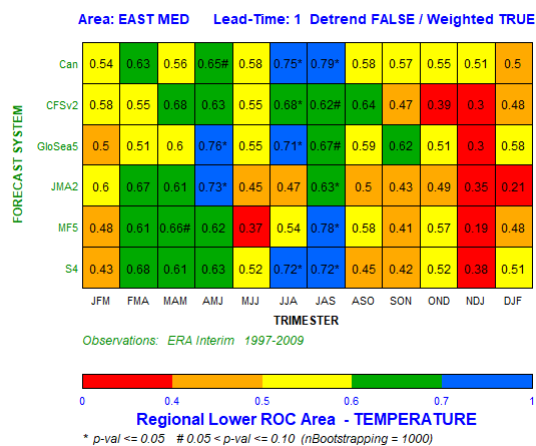
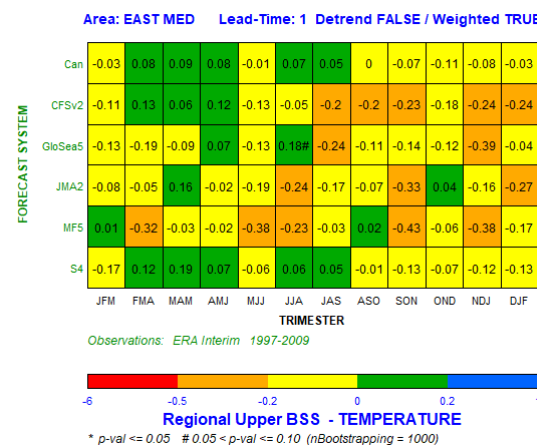
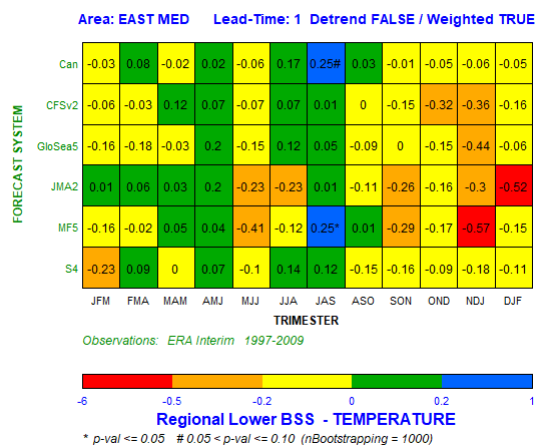
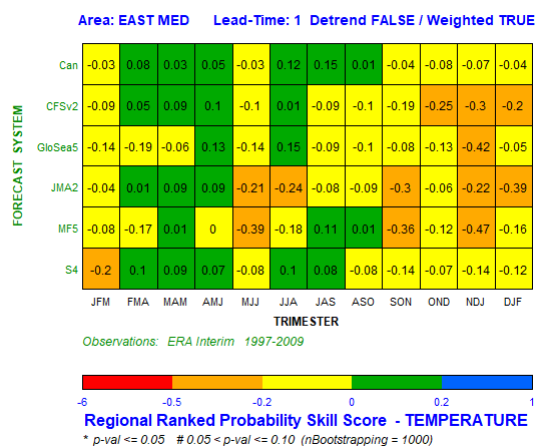
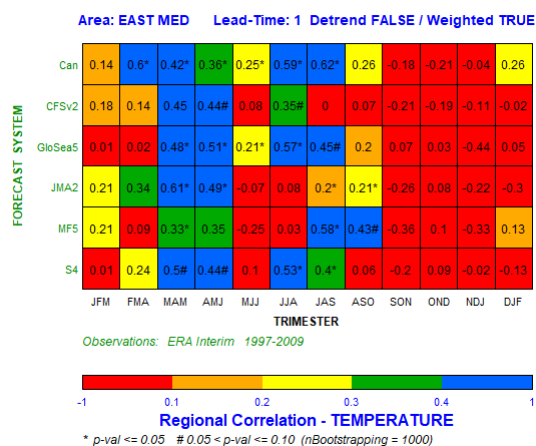


Table 9. The same as Table 4, but for EAST MEDITERRANEAN domain

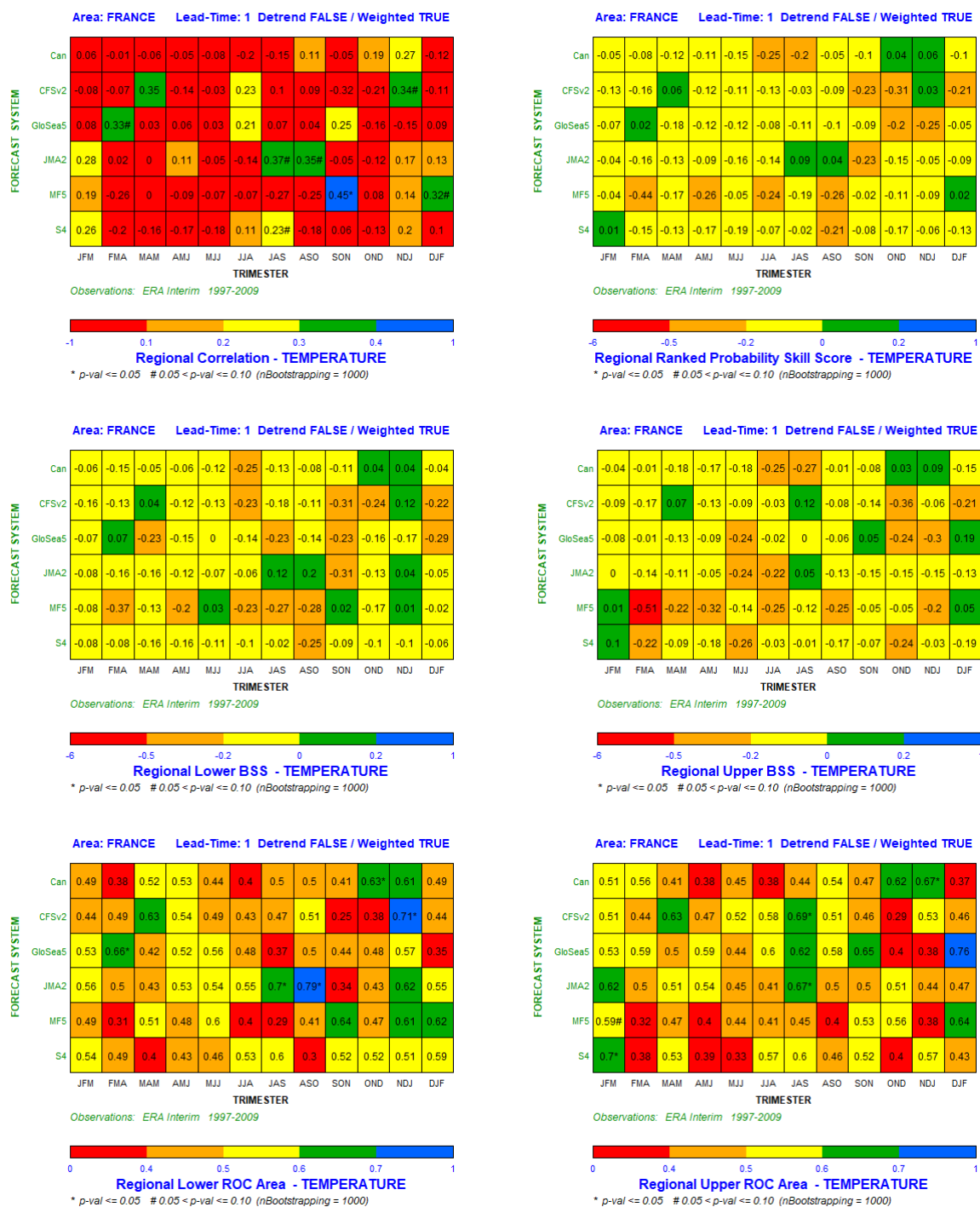


Table 10. The same as Table 4, but for FRANCE domain

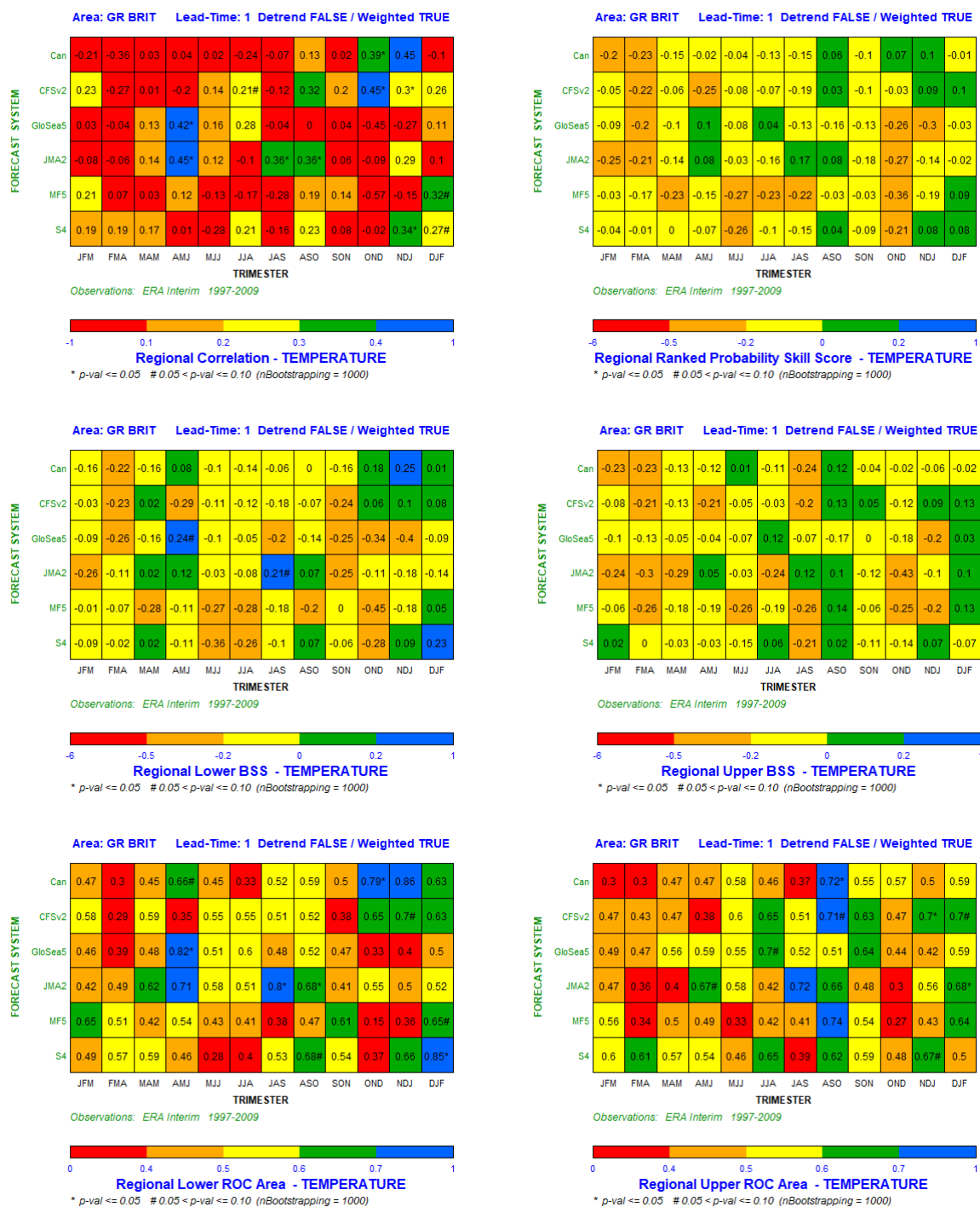


Table 11. The same as Table 4, but for GREAT BRITAIN domain

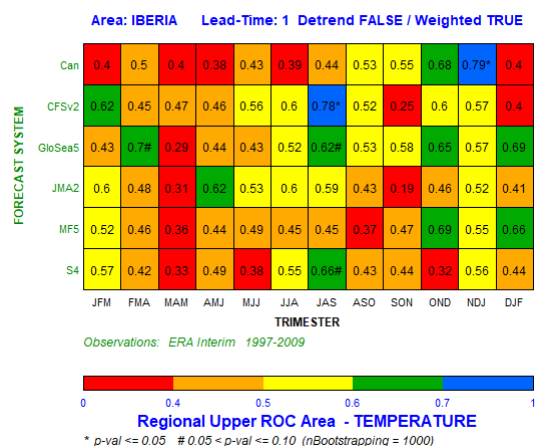
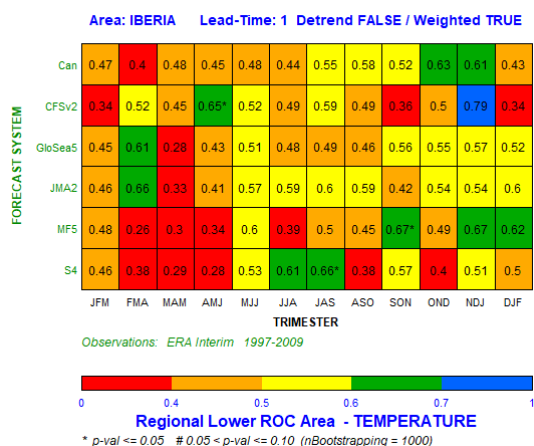
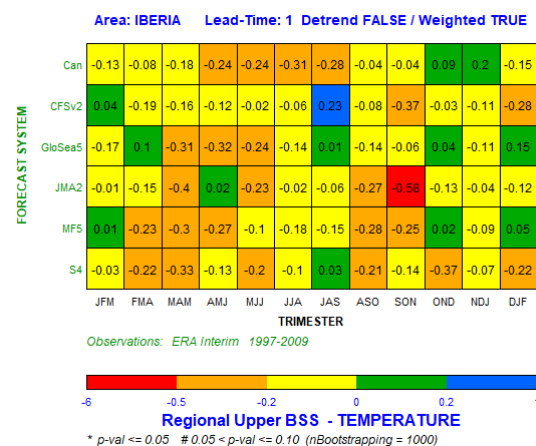
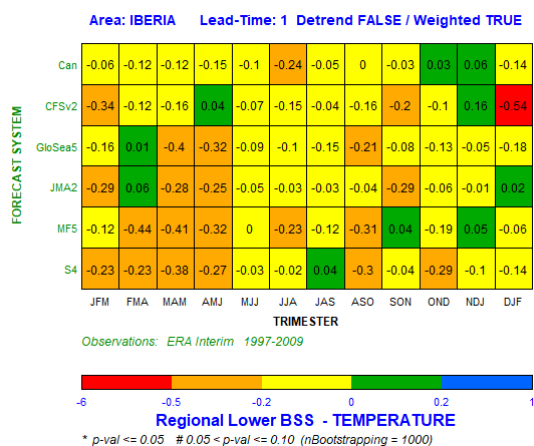
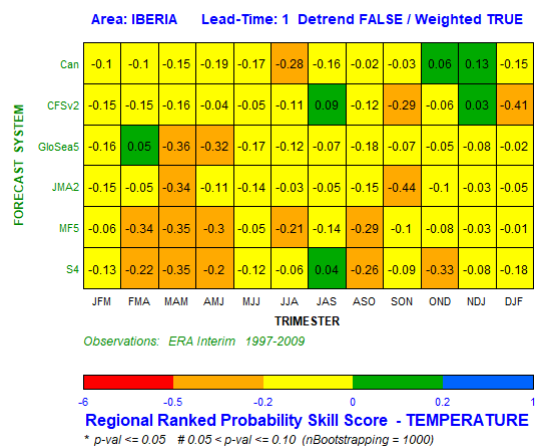
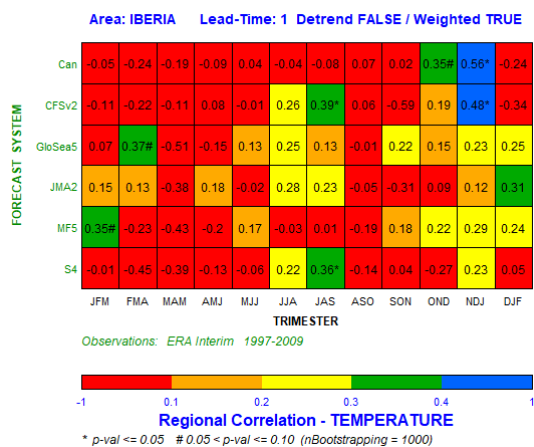


Table 12. The same as Table 4, but for IBERIA domain

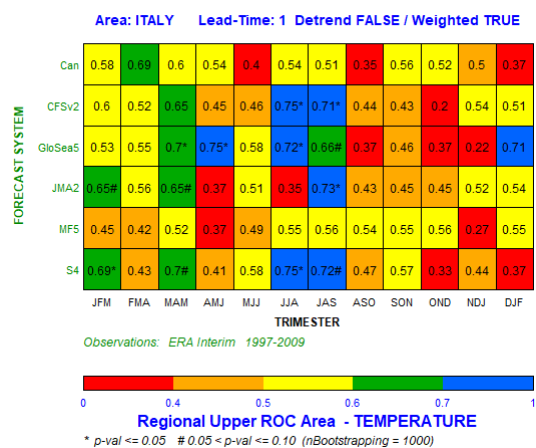
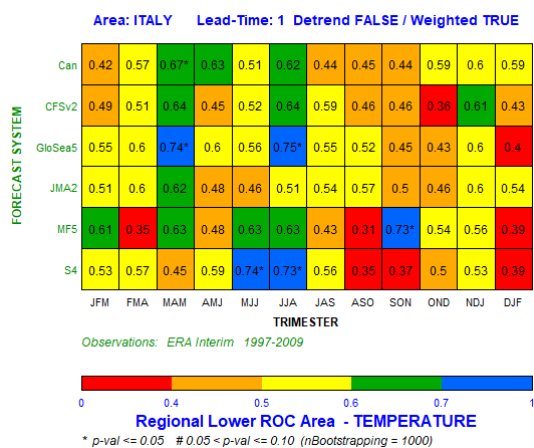
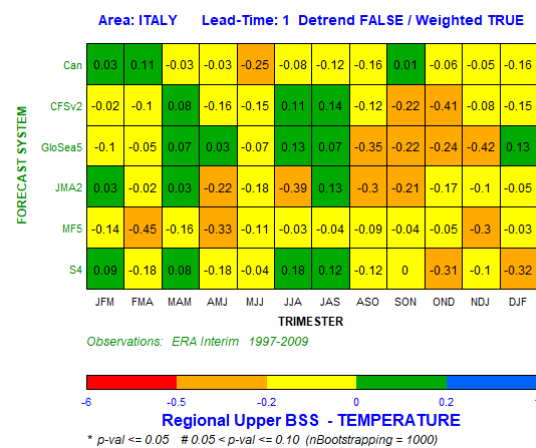
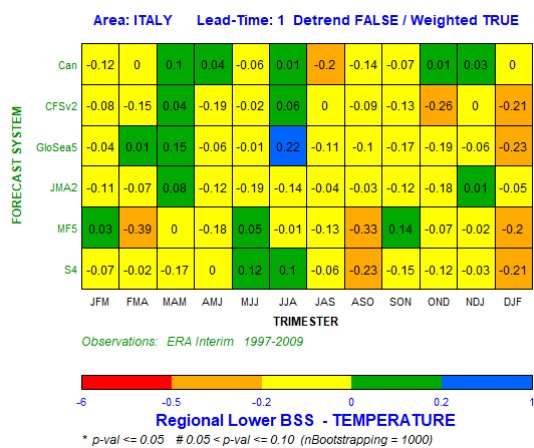
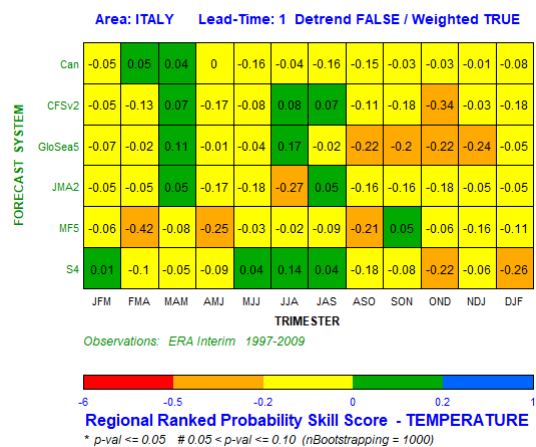
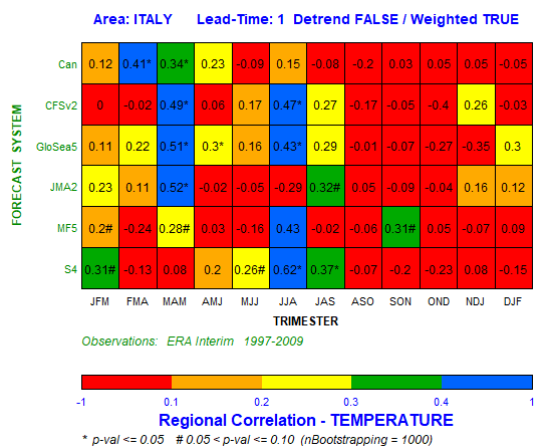


Table 13. The same as Table 4, but for ITALY domain

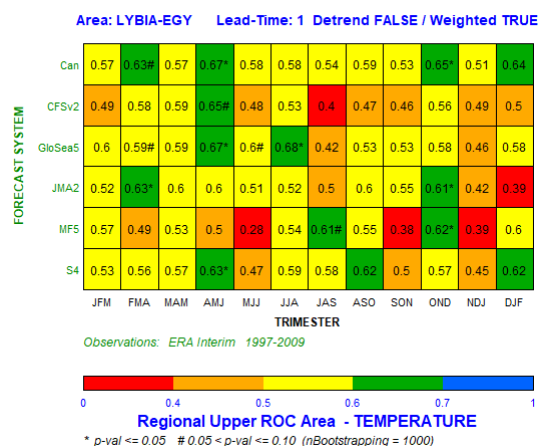
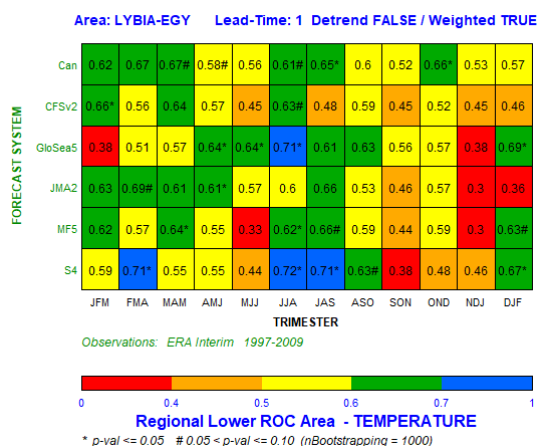
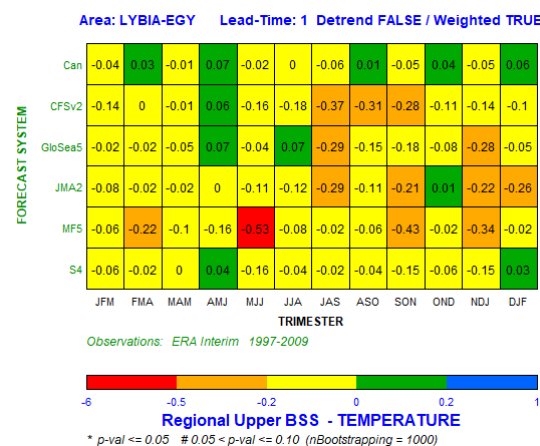
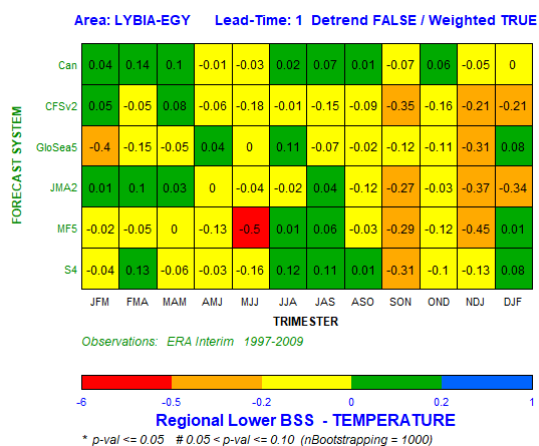
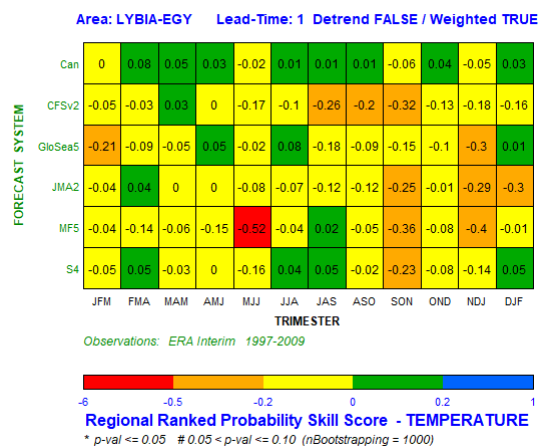
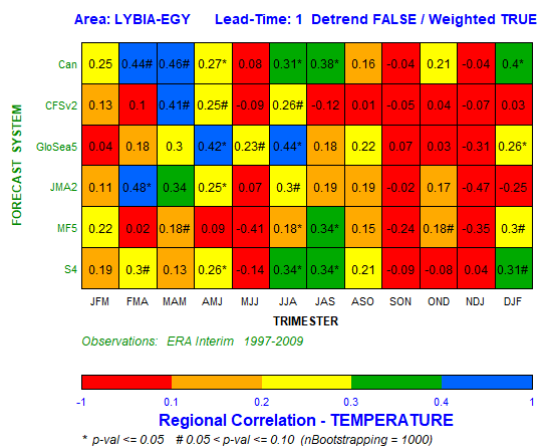


Table 14. The same as Table 4, but for LIBYA-EGYPT domain

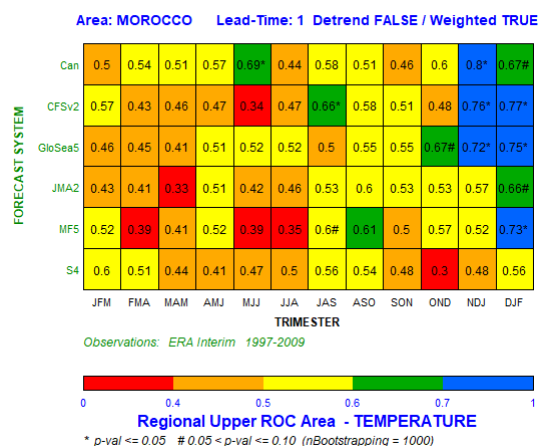
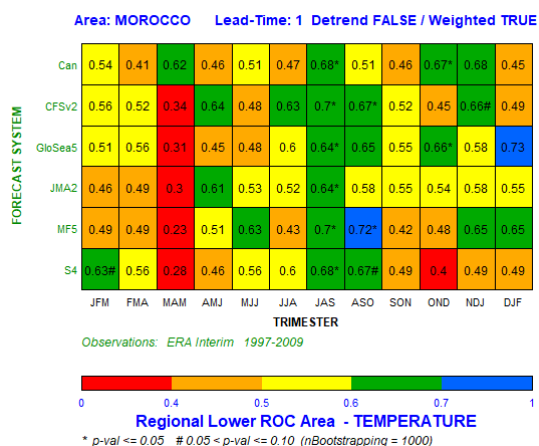
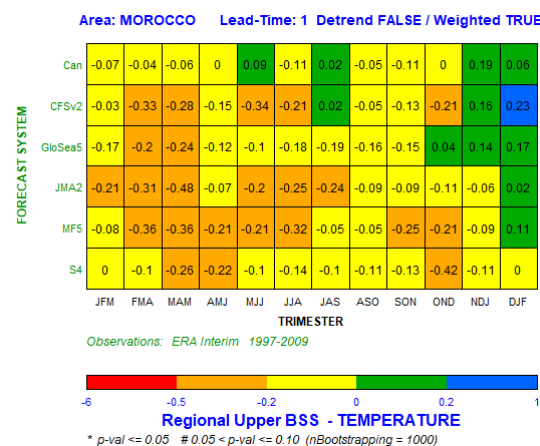
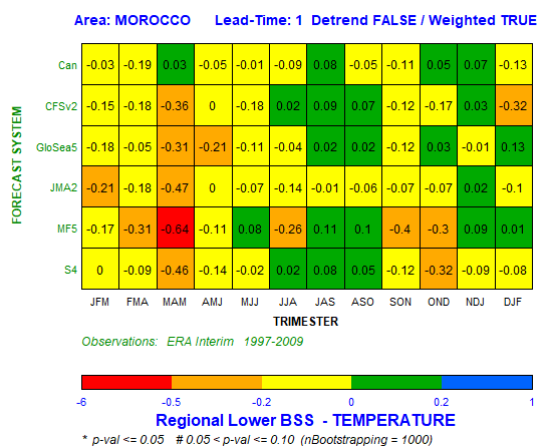
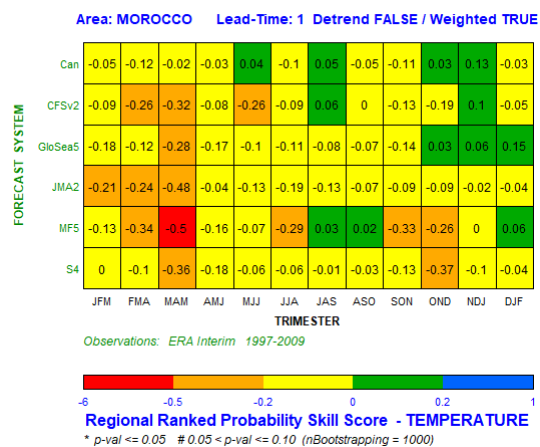
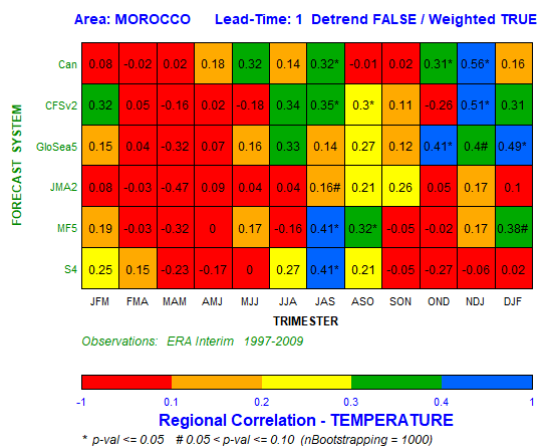


Table 15. The same as Table 4, but for MOROCCO domain

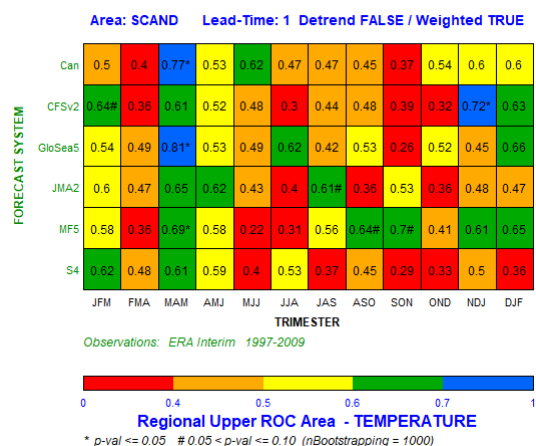
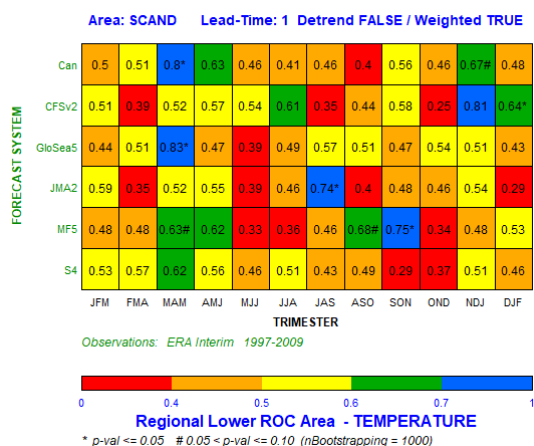
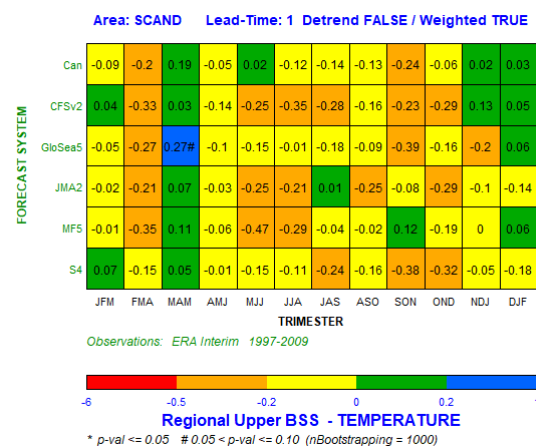
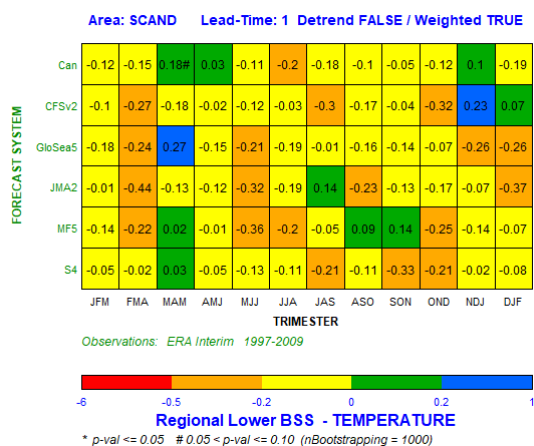
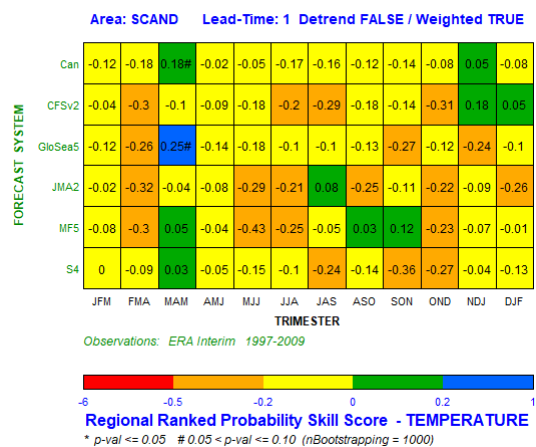
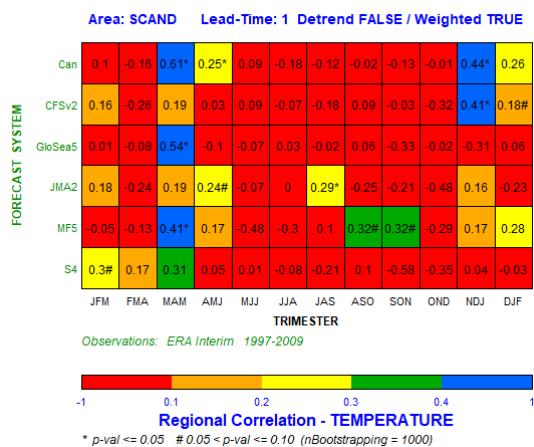


Table 16. The same as Table 4, but for SCANDINAVIA domain

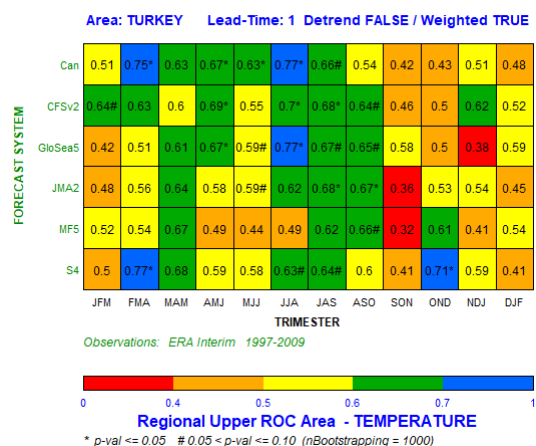
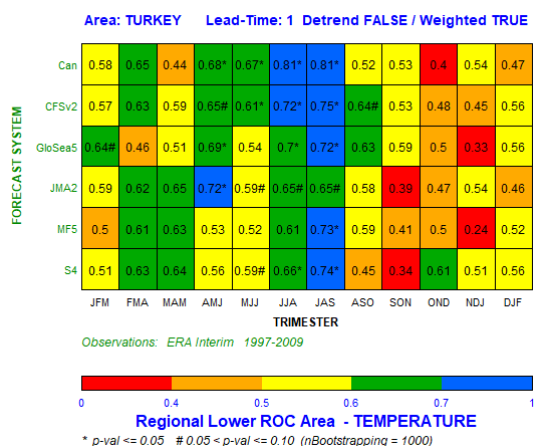
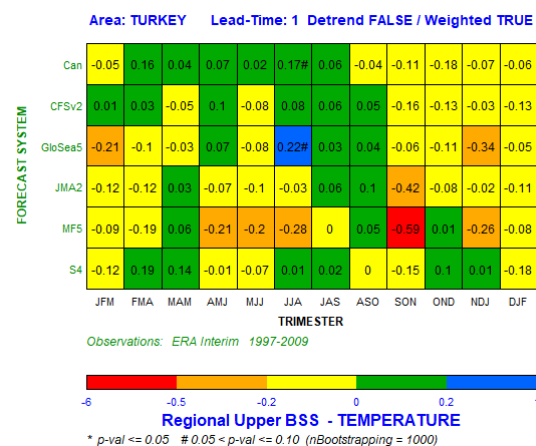
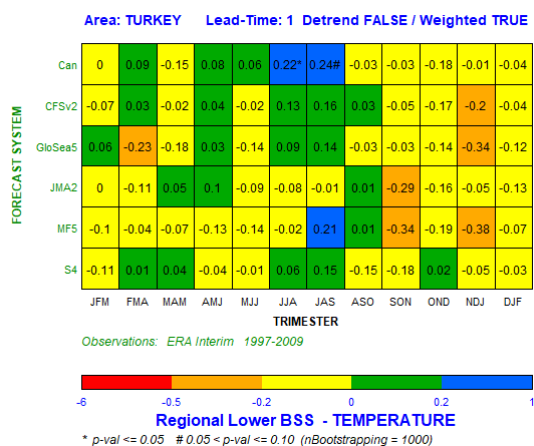
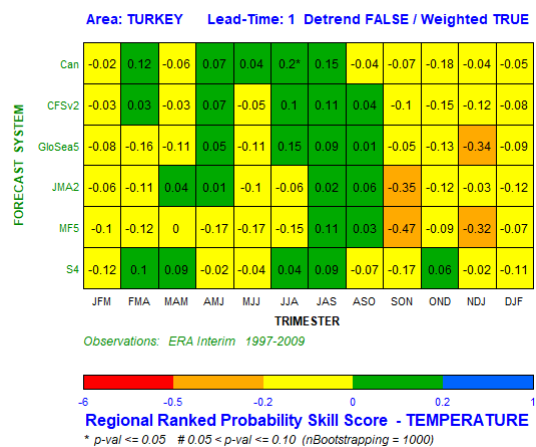
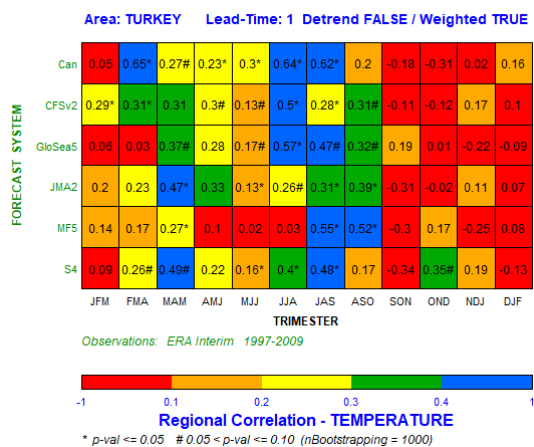


Table 17. The same as Table 4, but for TURKEY domain

4.1.b. Precipitation

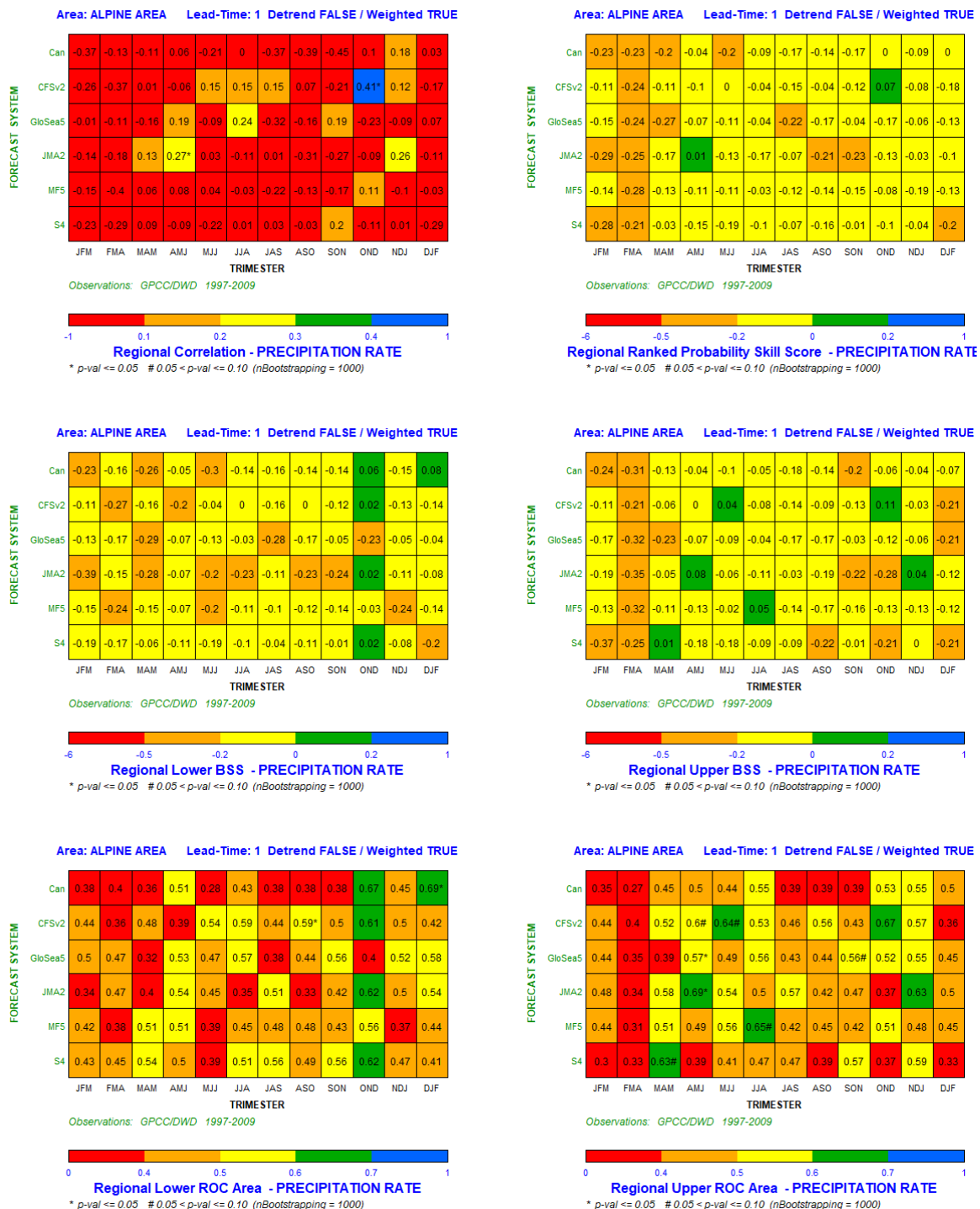


Table 18. Regional correlation coefficient, Ranked Probability Skill Score, Lower and Upper ROC Area and Brier Skill Score for precipitation anomalies, computed for 12 different three-month periods and lead-time 1 over ALPINE AREA domain (verification period: 1997-2009). Different three-month periods are shown in the X-axis and systems (S4, MF5, JMA2, GLOSEA5, CFSv2 and Can) are represented in the Y-axis (see text for description).

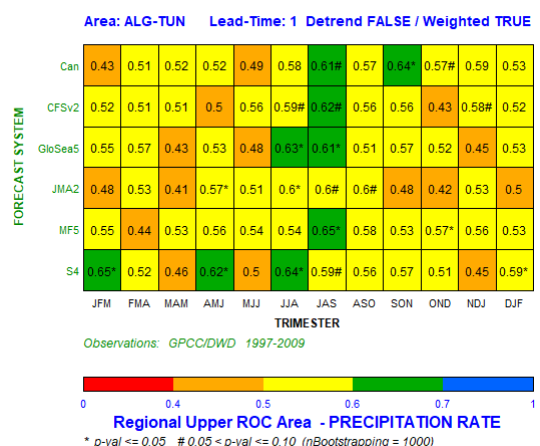
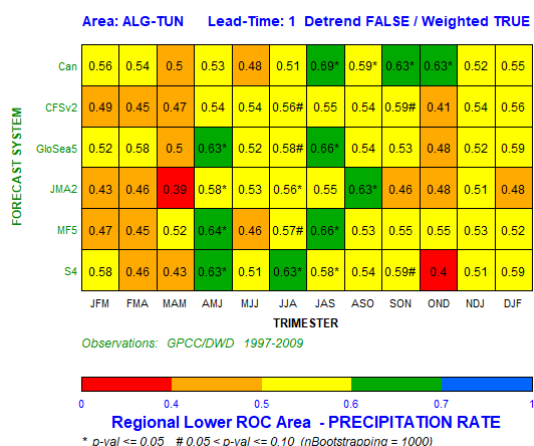
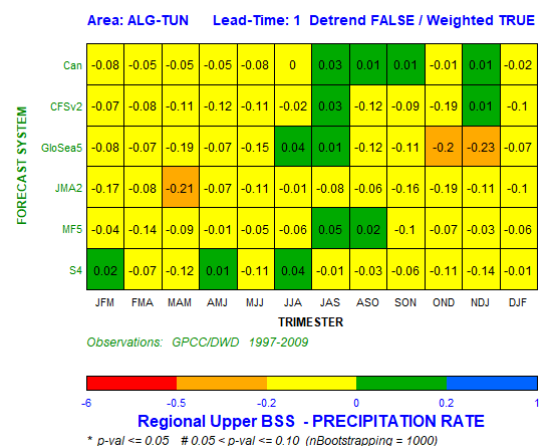
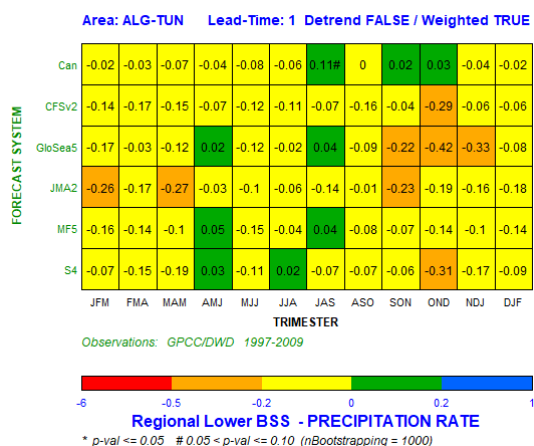
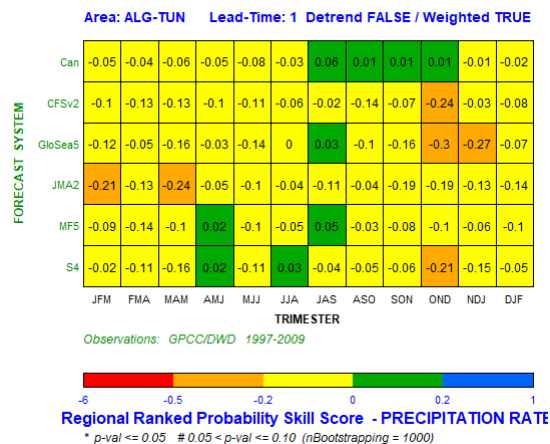
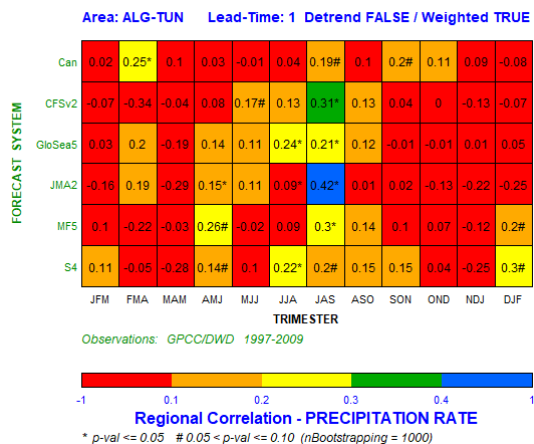


Table 19. The same as Table 18, but for ALGERIA/TUNISIA domain

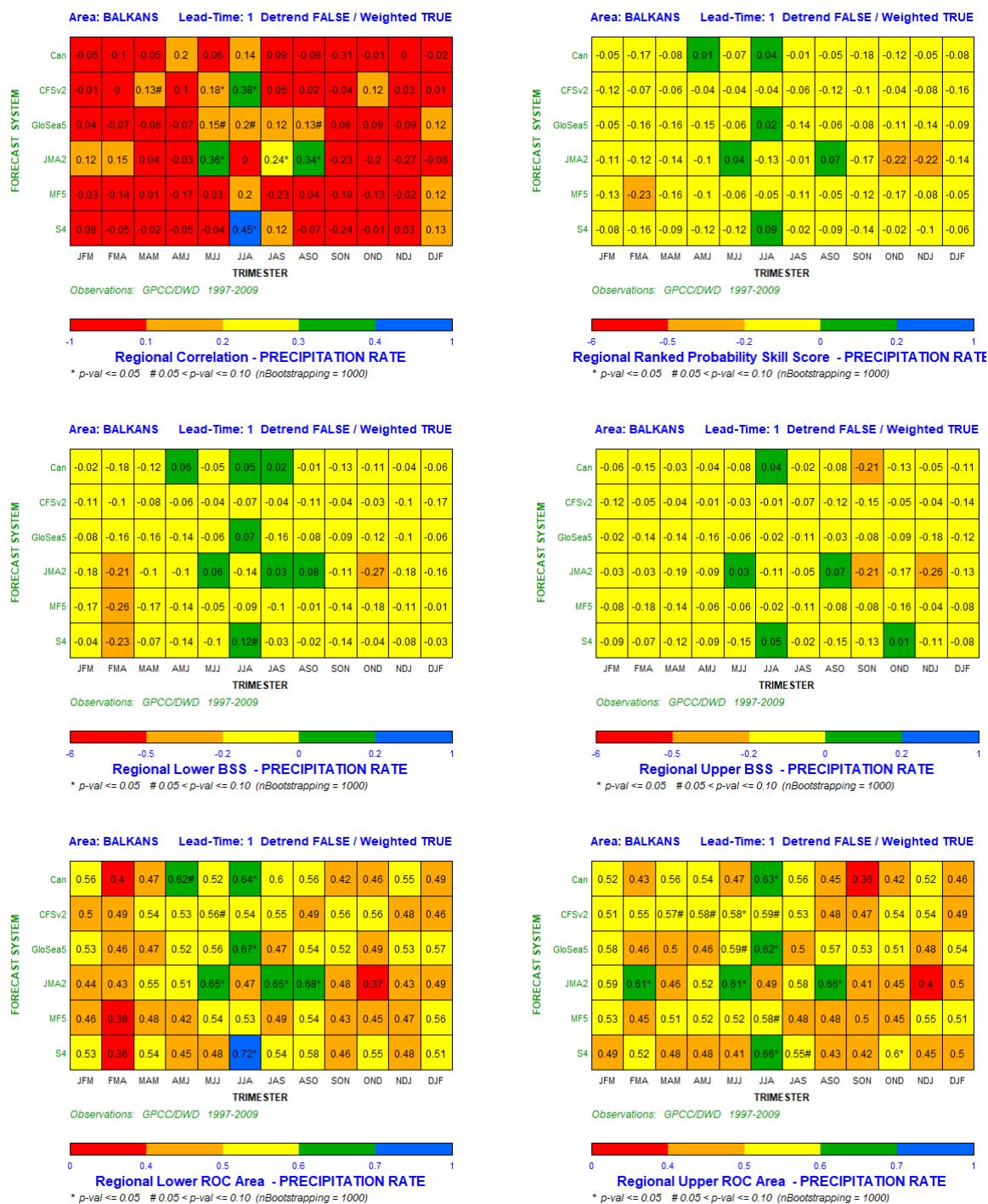


Table 20. The same as Table 18, but for BALKANS domain

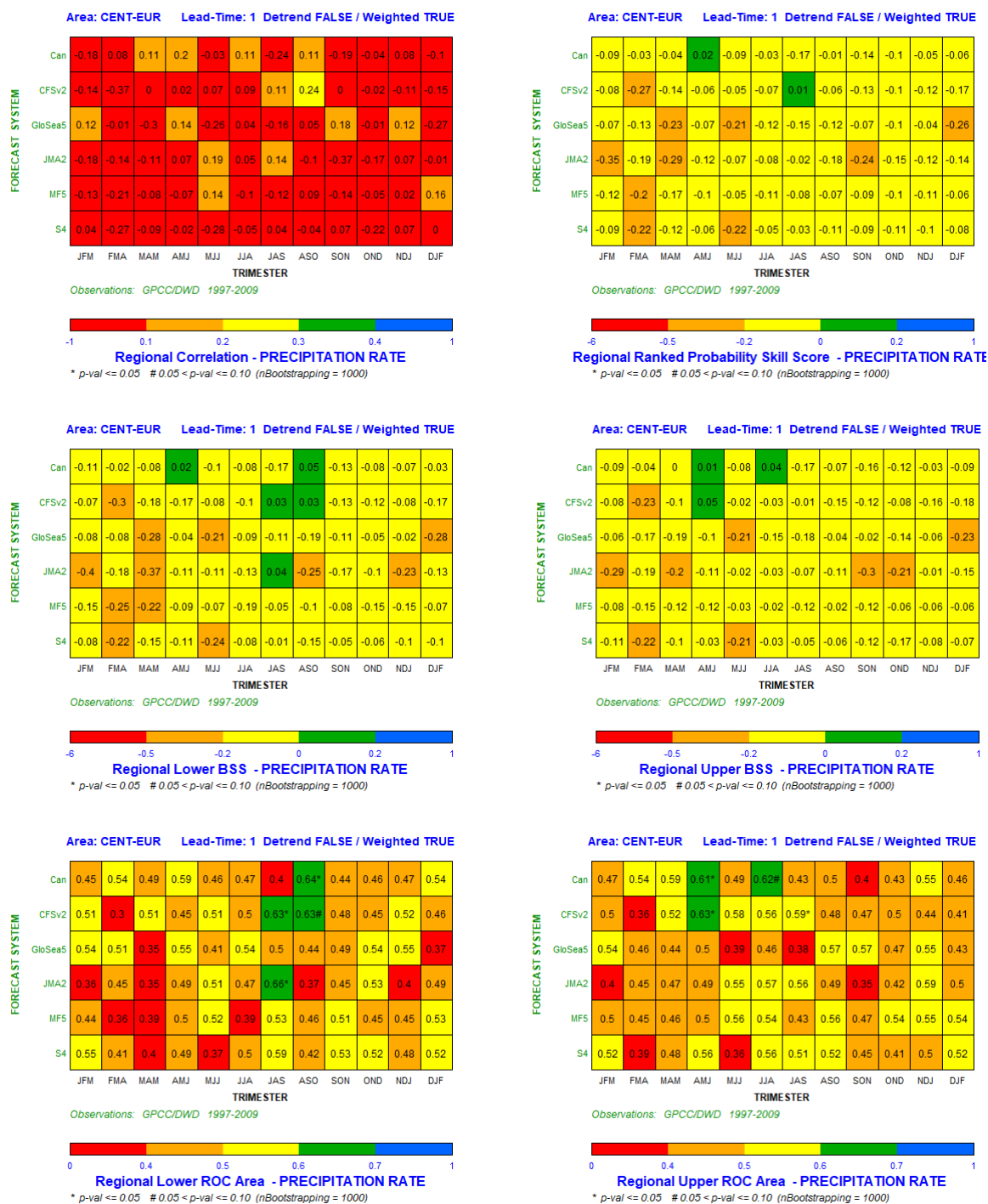


Table 21. The same as Table 18, but for CENTRE EUROPE domain

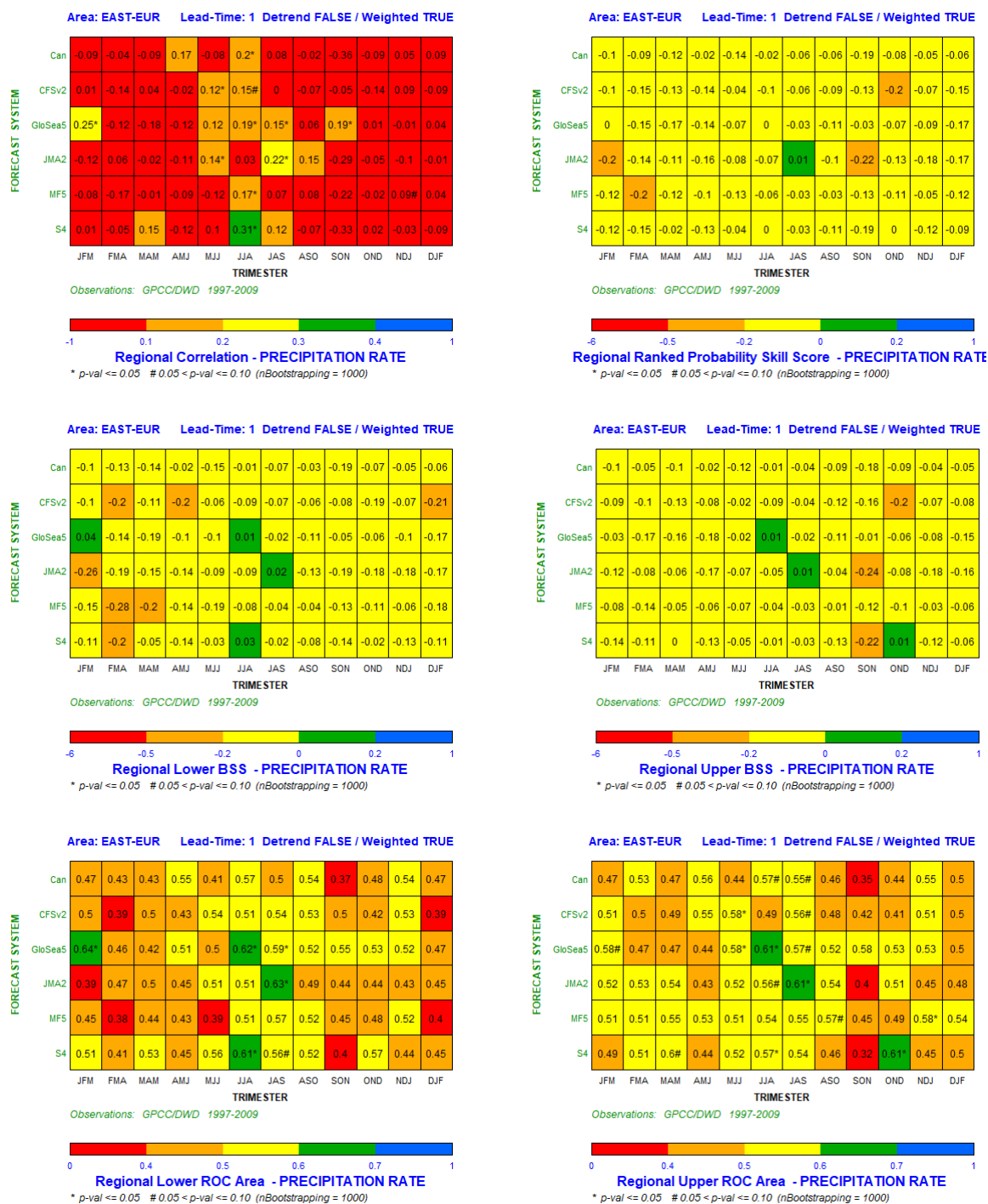


Table 22. The same as Table 18, but for EAST EUROPE domain

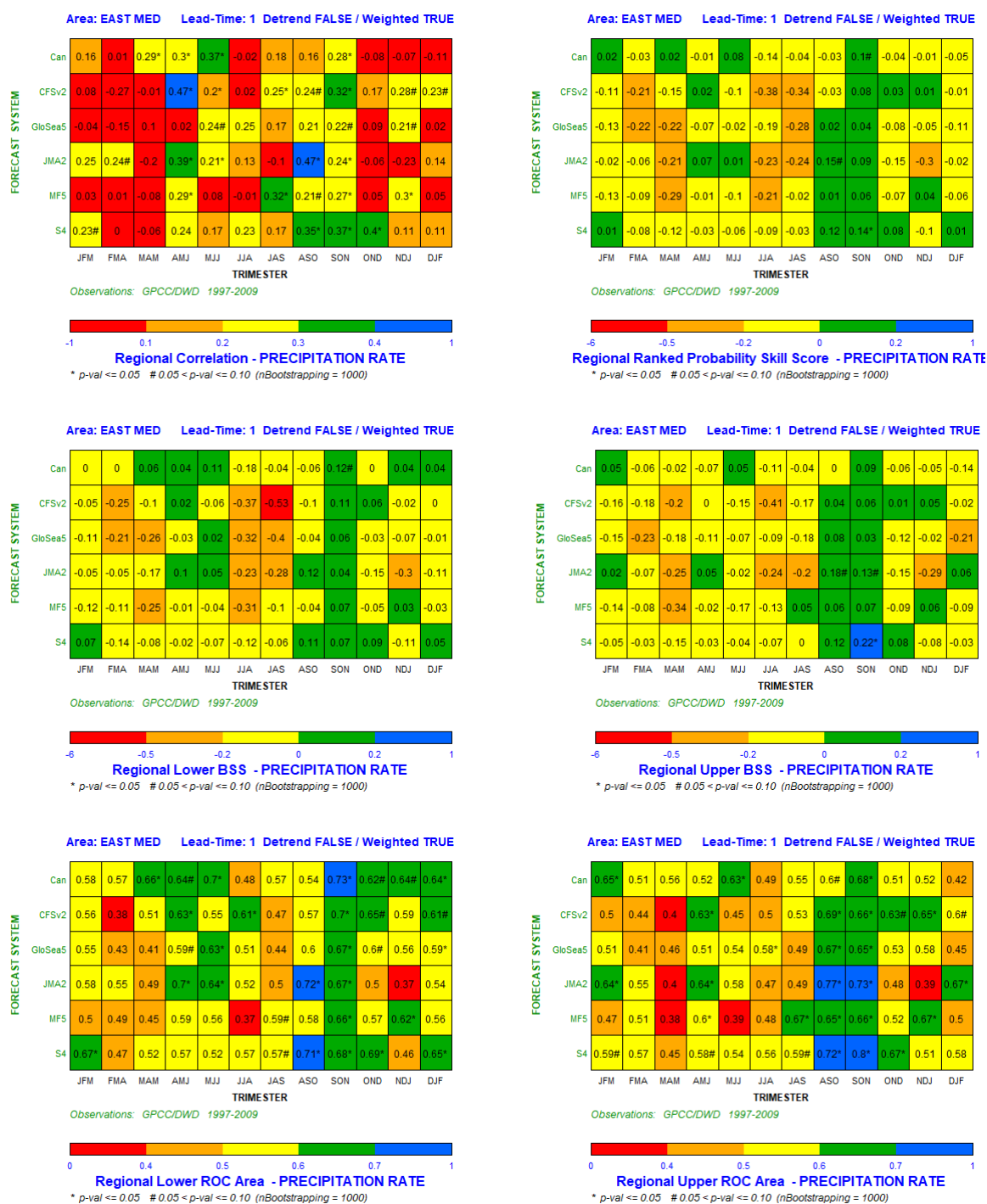


Table 23. The same as Table 18, but for EAST MEDITERRANEAN domain

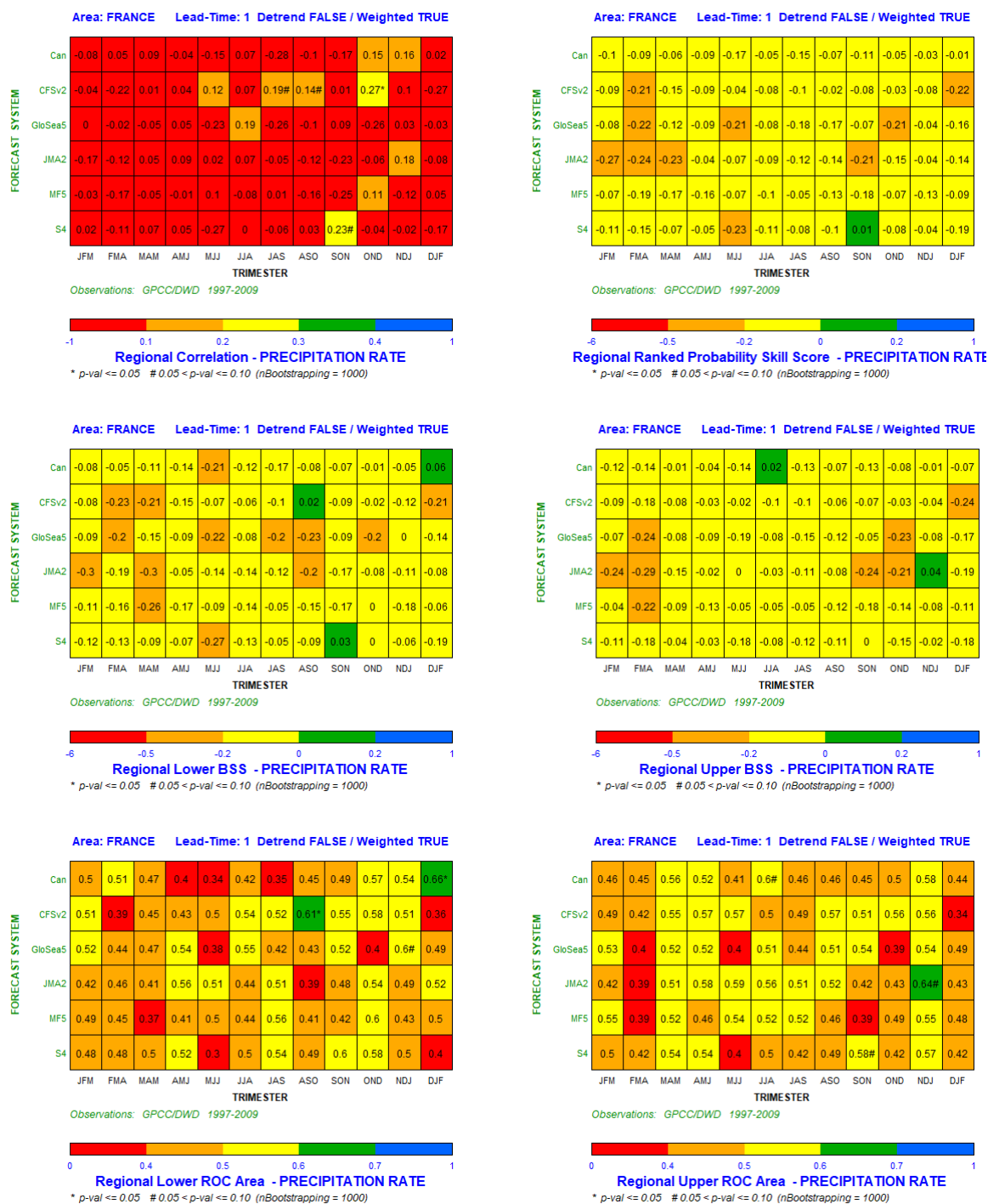


Table 24. The same as Table 18, but for FRANCE domain

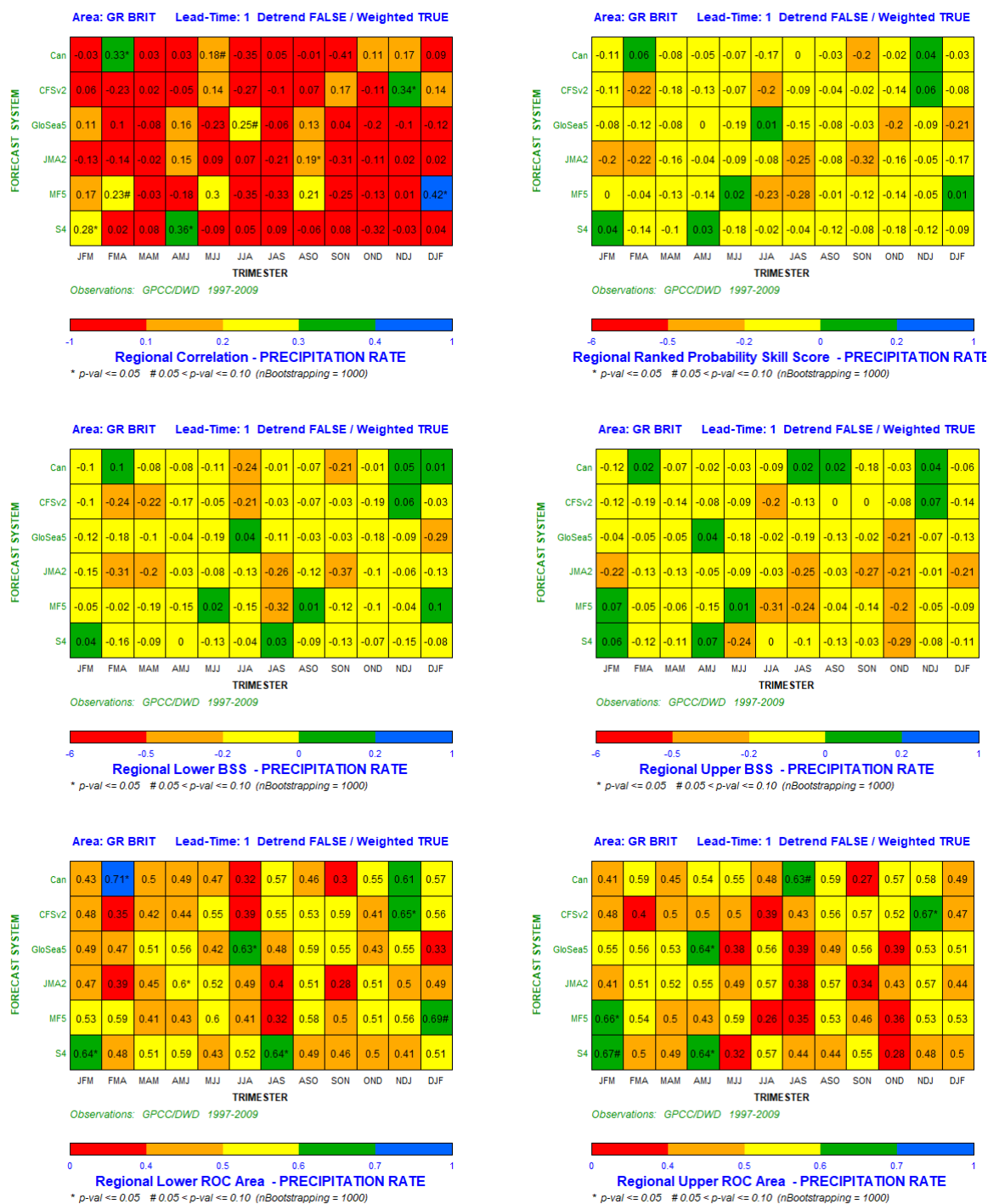


Table 25. The same as Table 18, but for GREAT BRITAIN domain

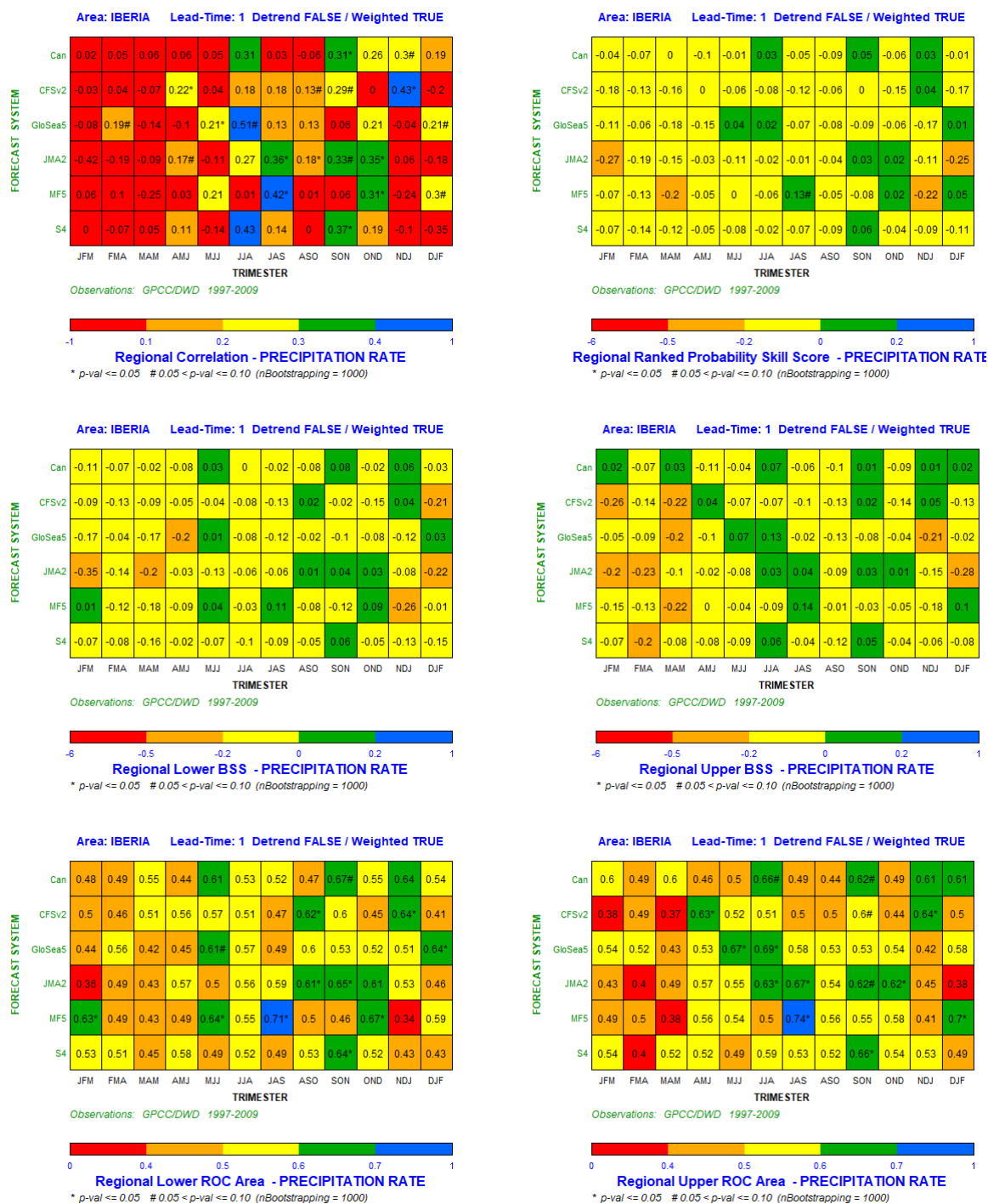


Table 26. The same as Table 18, but for IBERIA domain

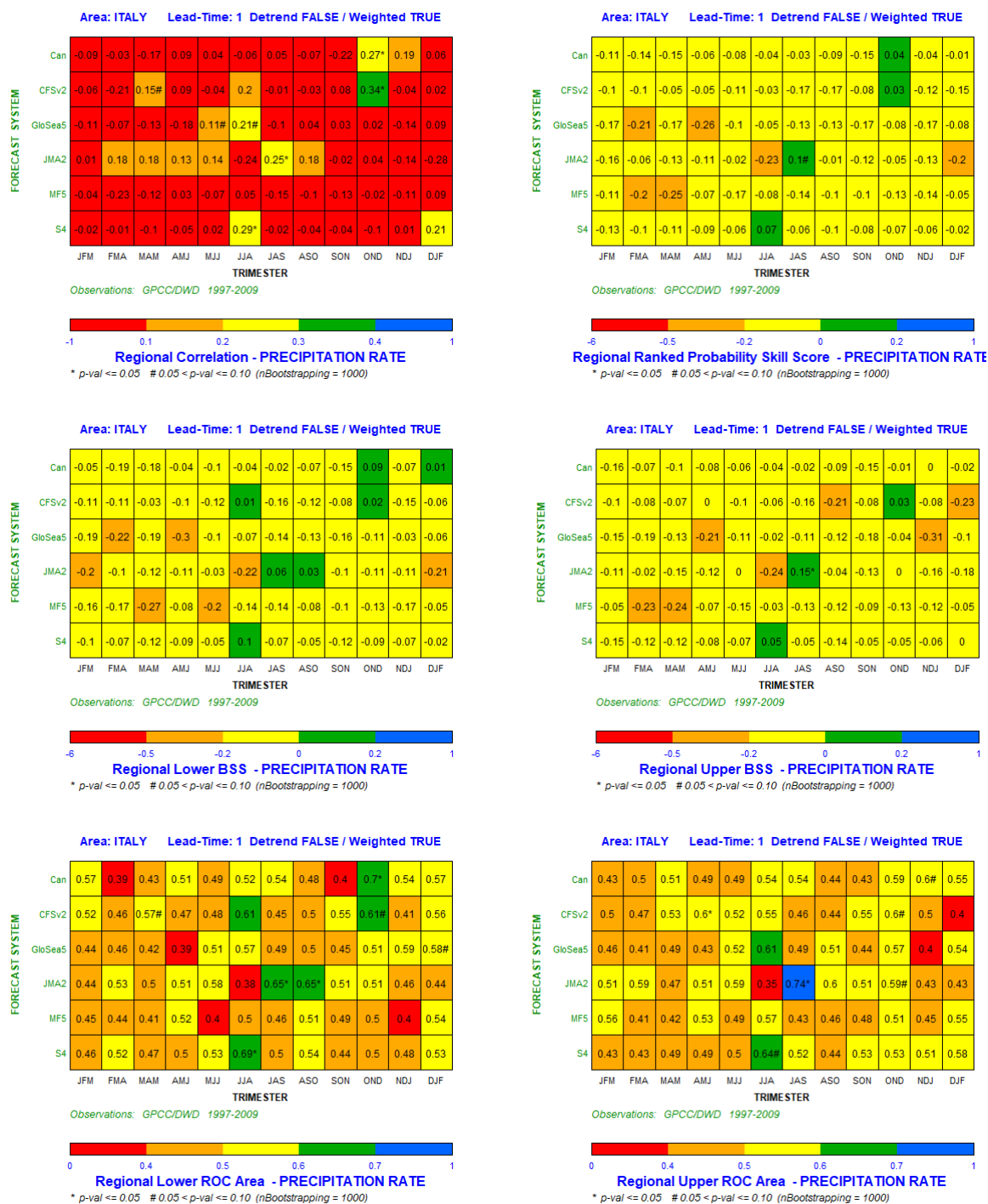


Table 27. The same as Table 18, but for ITALY domain

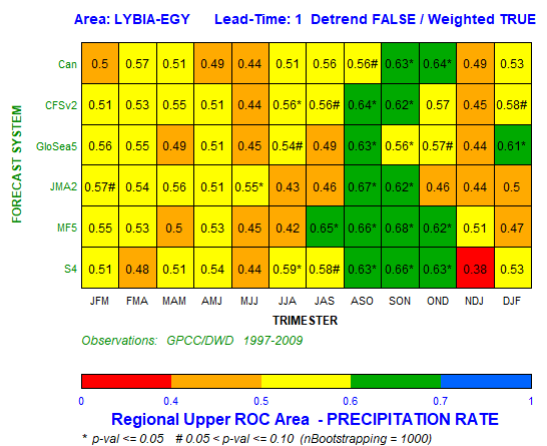
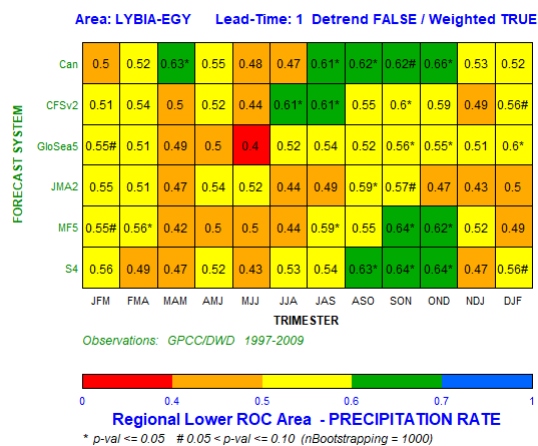
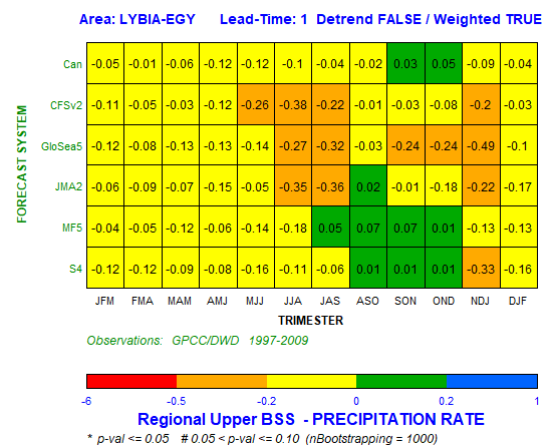
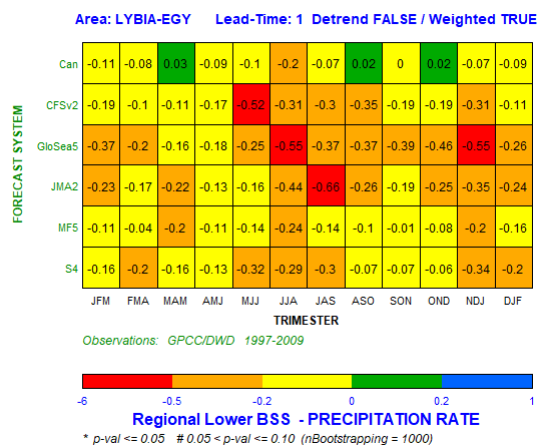
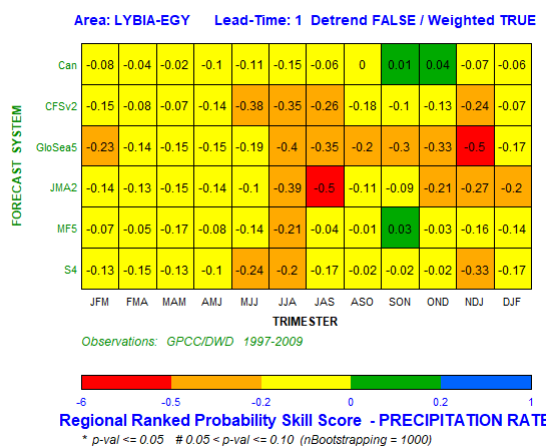
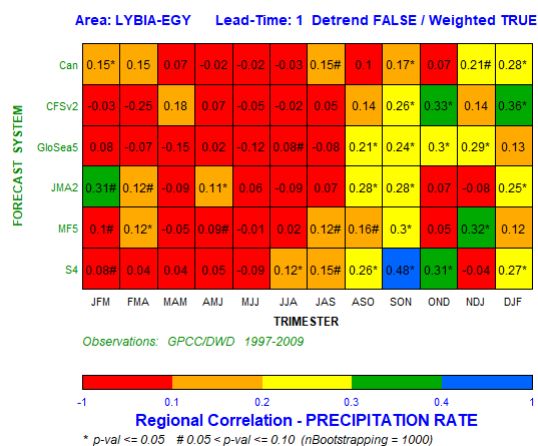


Table 28. The same as Table 18, but for LIBYA-EGYPT domain

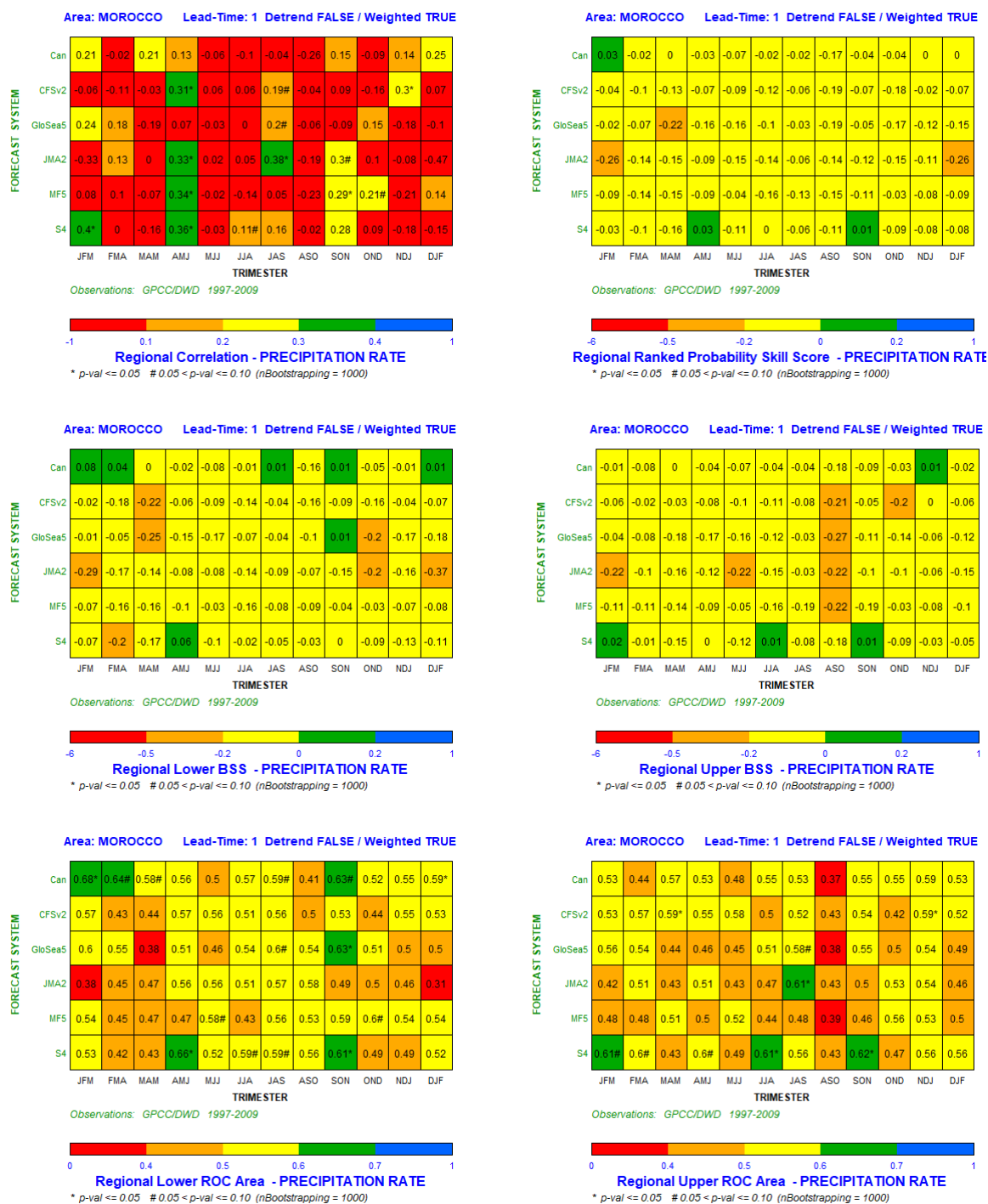


Table 29. The same as Table 18, but for MOROCCO domain

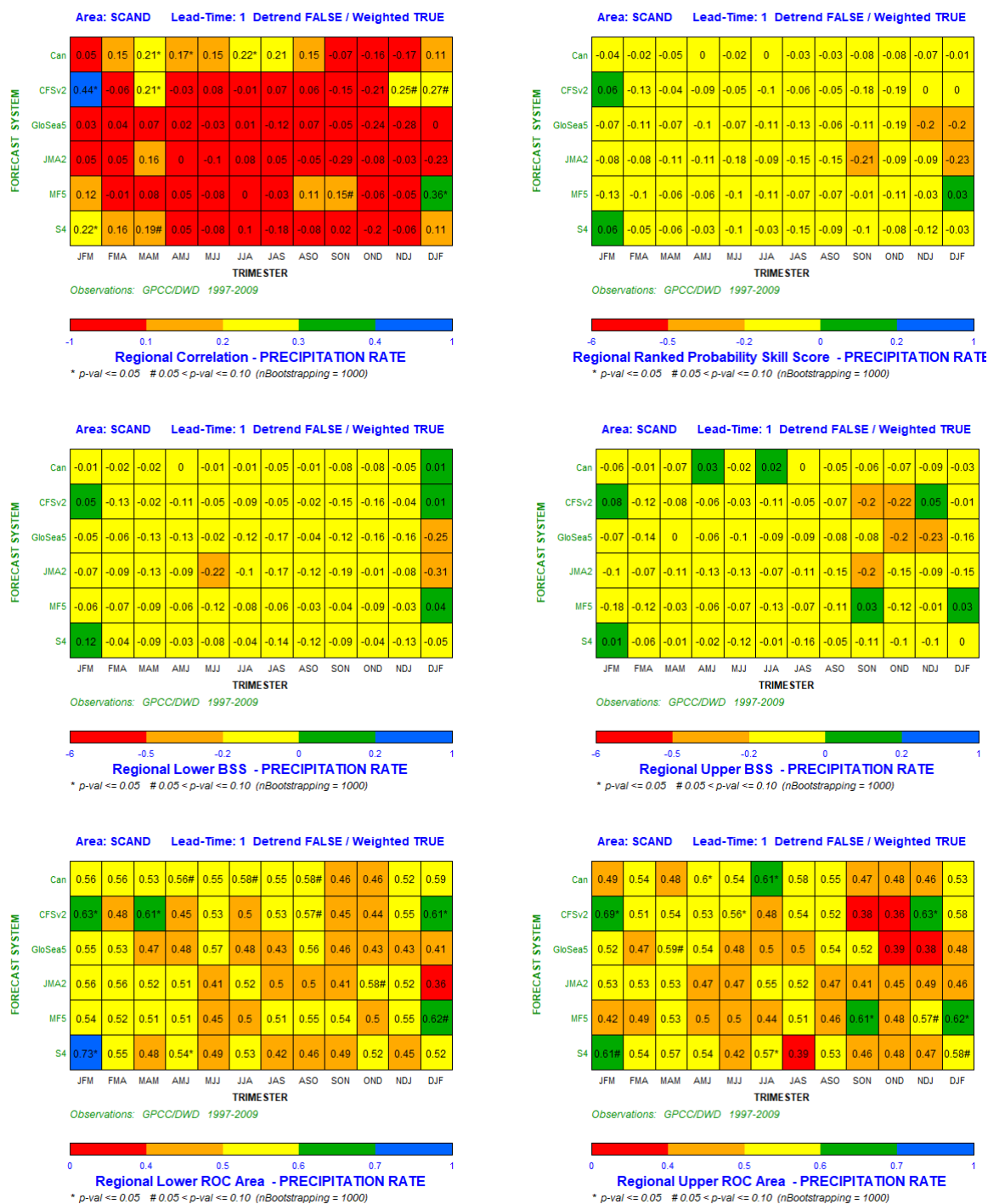


Table 30. The same as Table 18, but for SCANDINAVIA domain

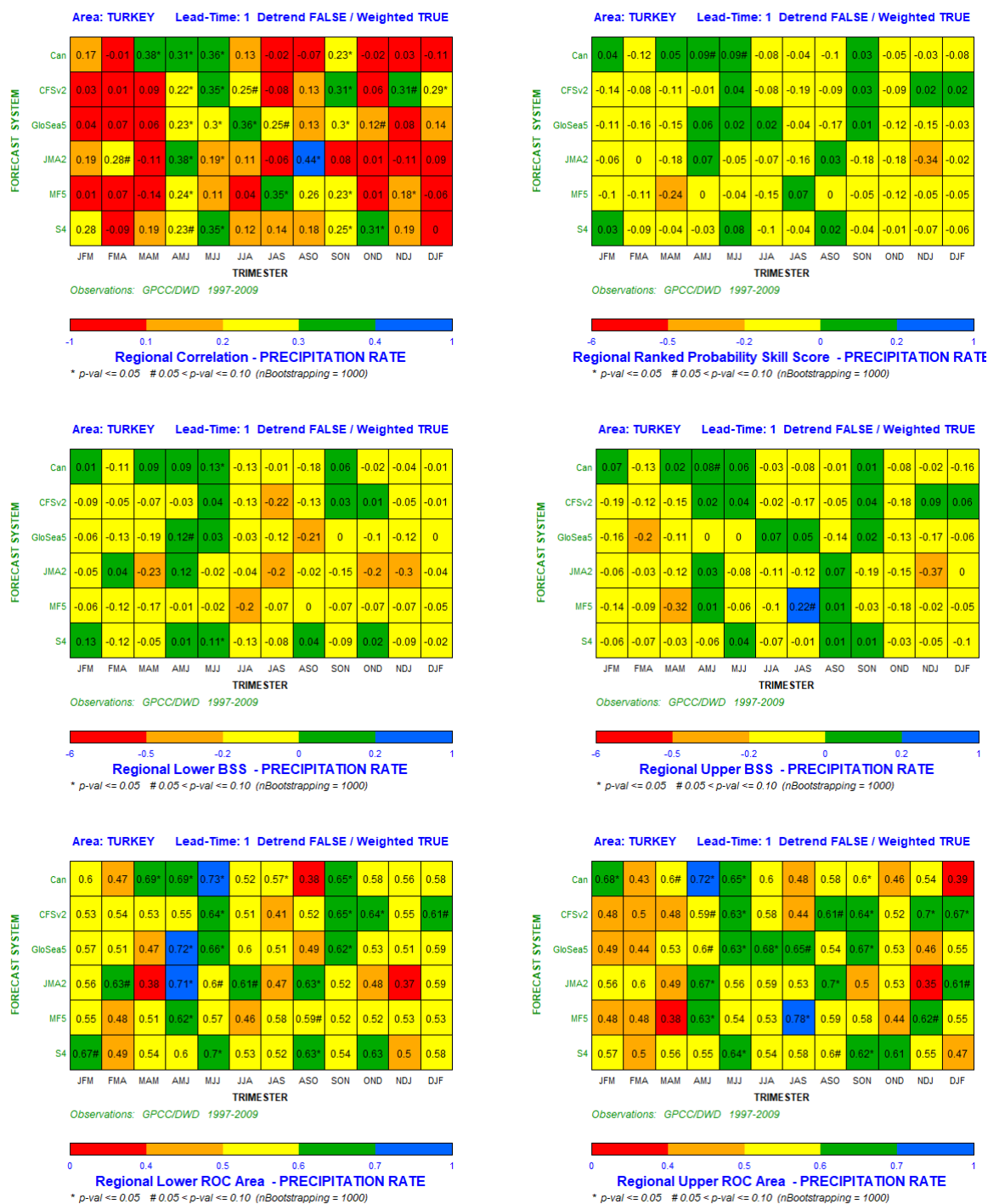


Table 31. The same as Table 18, but for TURKEY domain

4.2. Verification period: the maximum available period for each model

4.2.a. Temperature

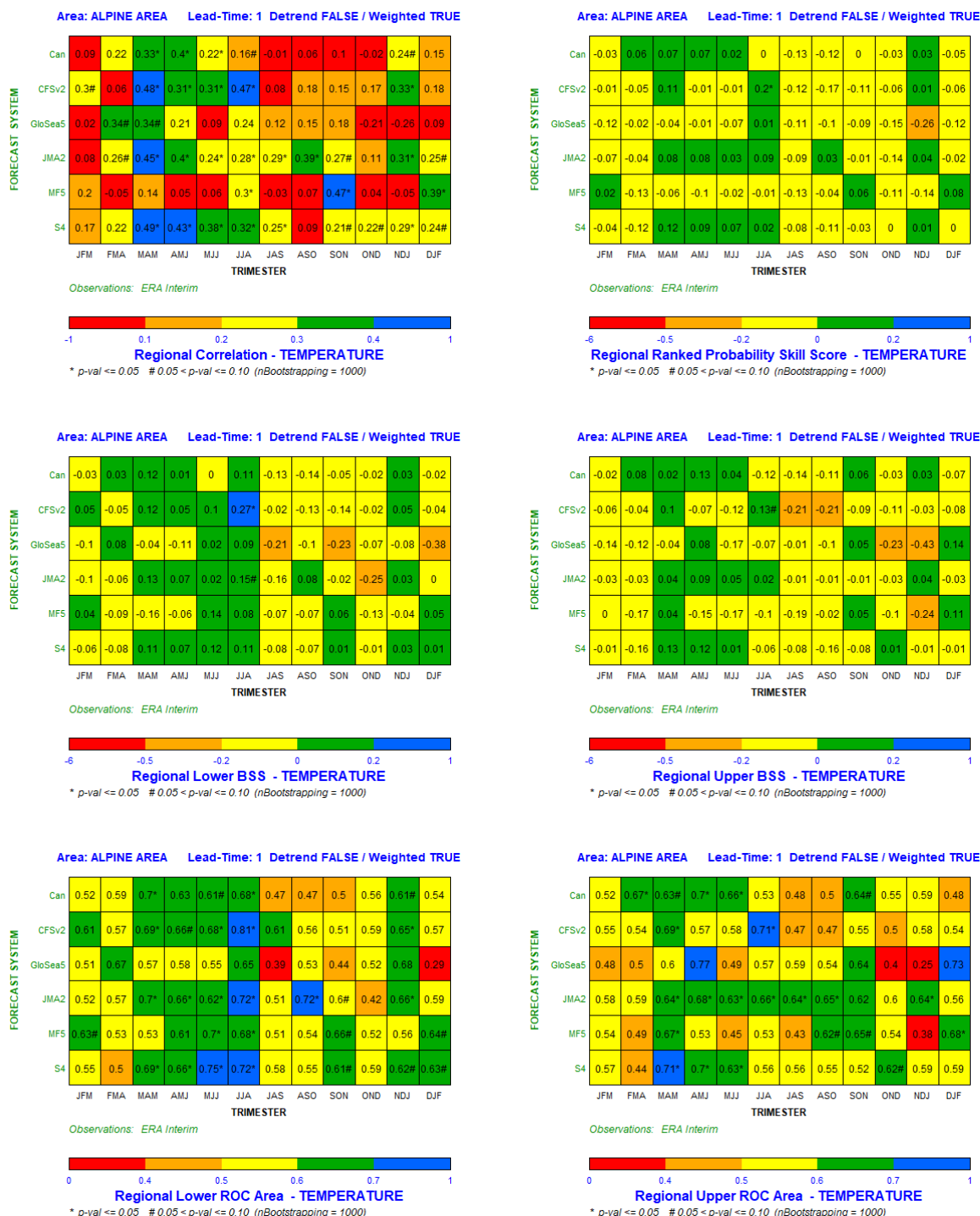


Table 32. Regional correlation coefficient, Ranked Probability Skill Score, Lower and Upper ROC Area and Brier Skill Score computed for temperature anomalies (without removing trend), for 12 different three-month periods and lead-time 1 over ALPINE AREA domain (maximum available verification period for each system (see Table 1)). Three-month periods are shown in the X-axis and systems (S4, MF5, JMA2, GLOSEA5, CFSv2 and Can) are represented in the Y-axis (see text for description).

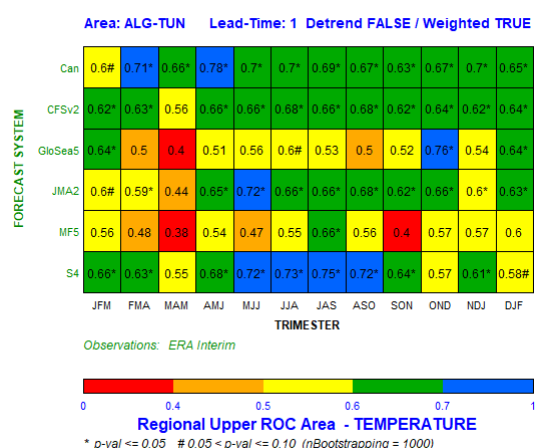
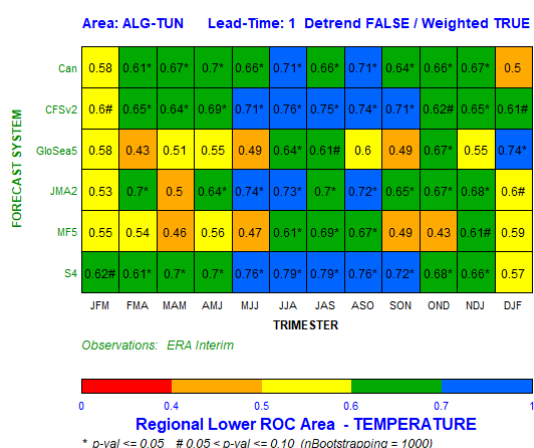
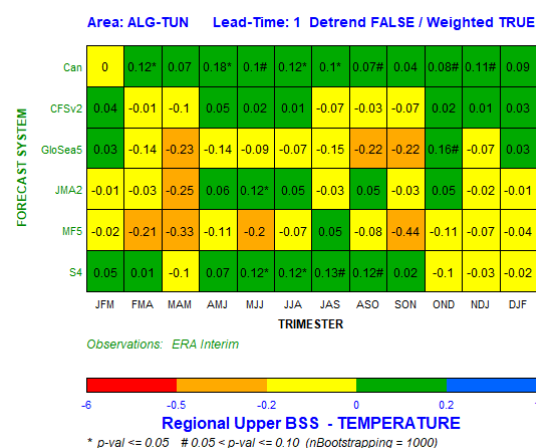
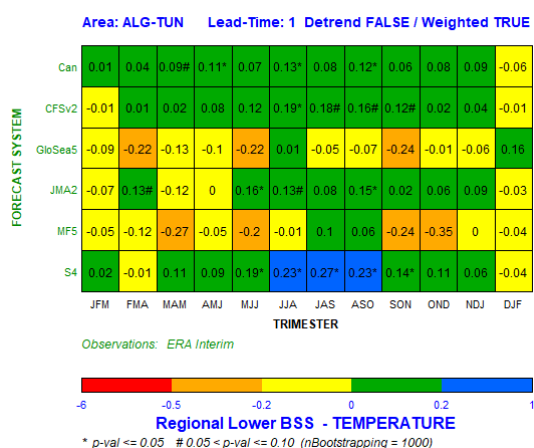
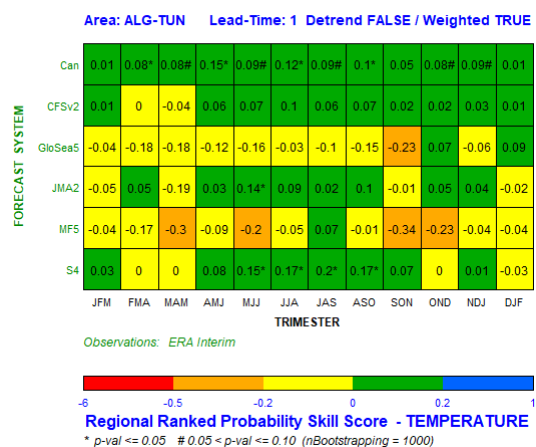
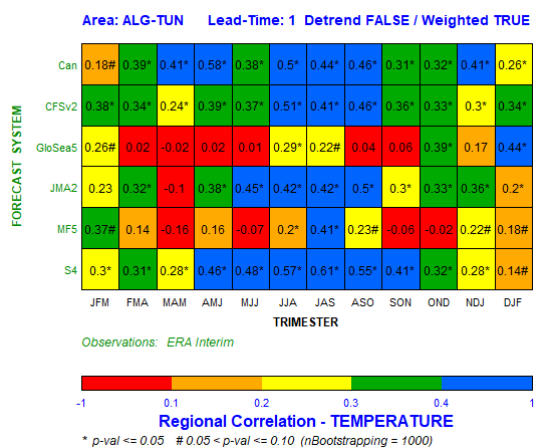


Table 33. The same as Table 32, but for ALGERIA/TUNISIA domain

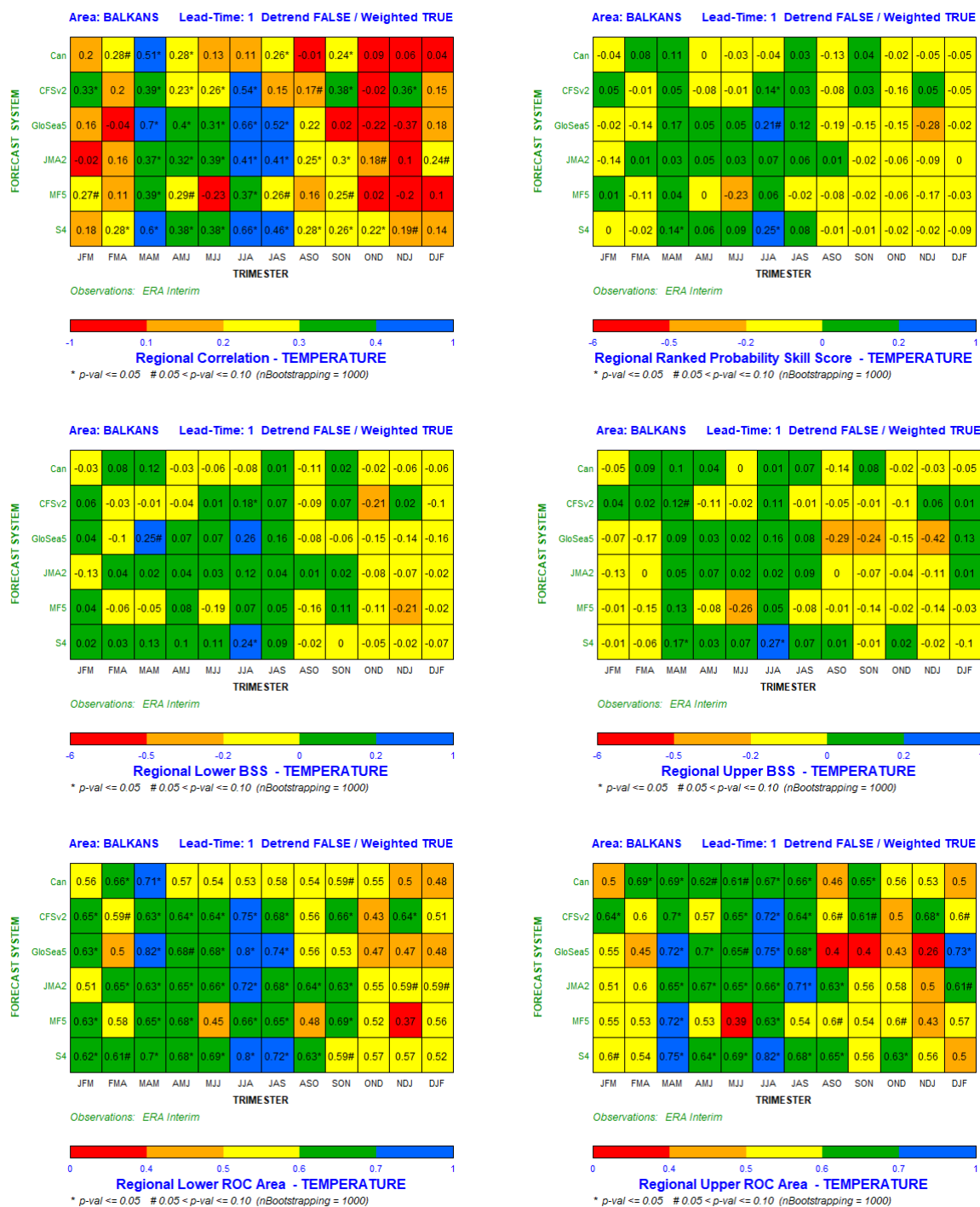


Table 34. The same as Table 32, but for BALKANS domain

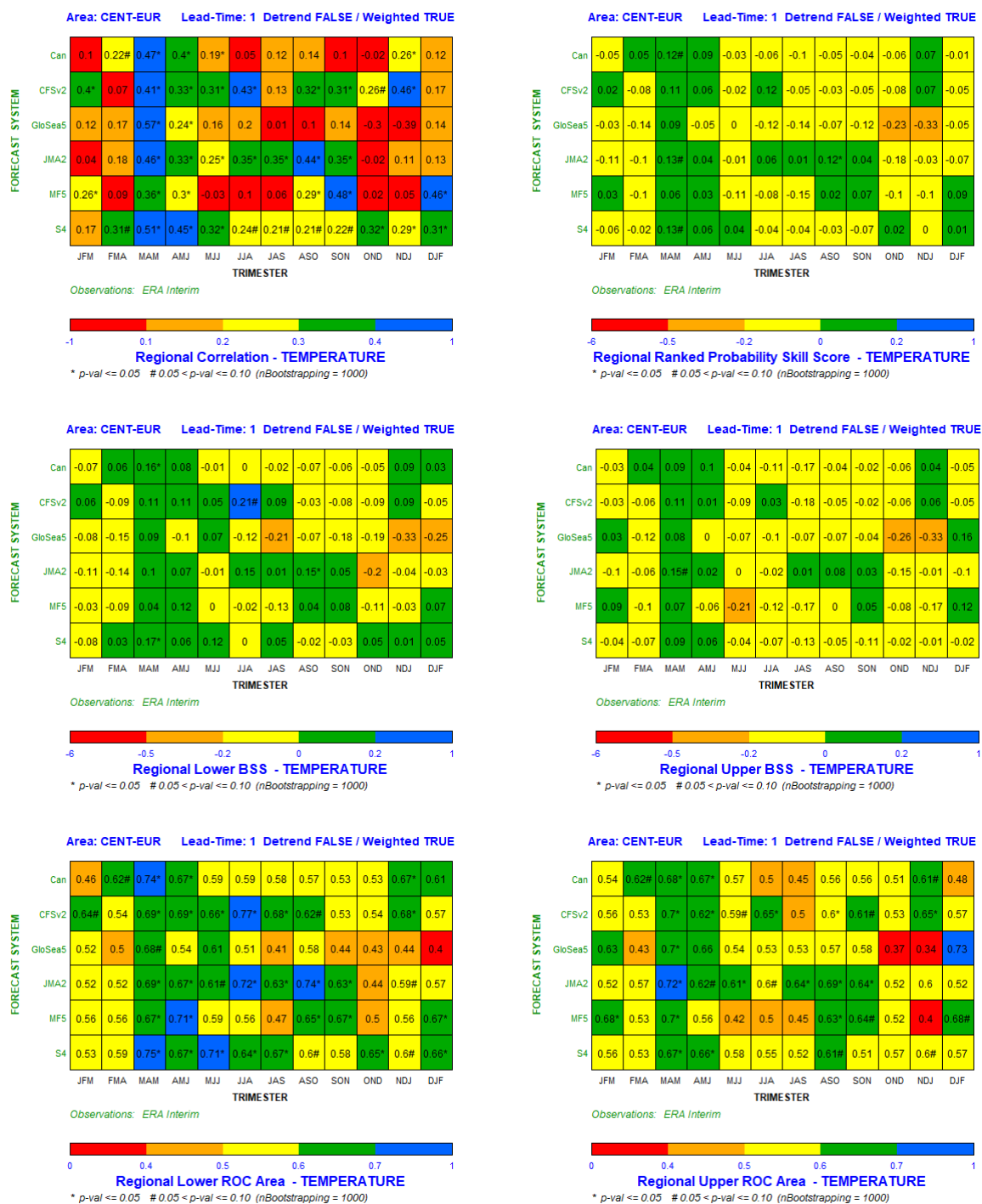


Table 35. The same as Table 32, but for CENTRE EUROPE domain

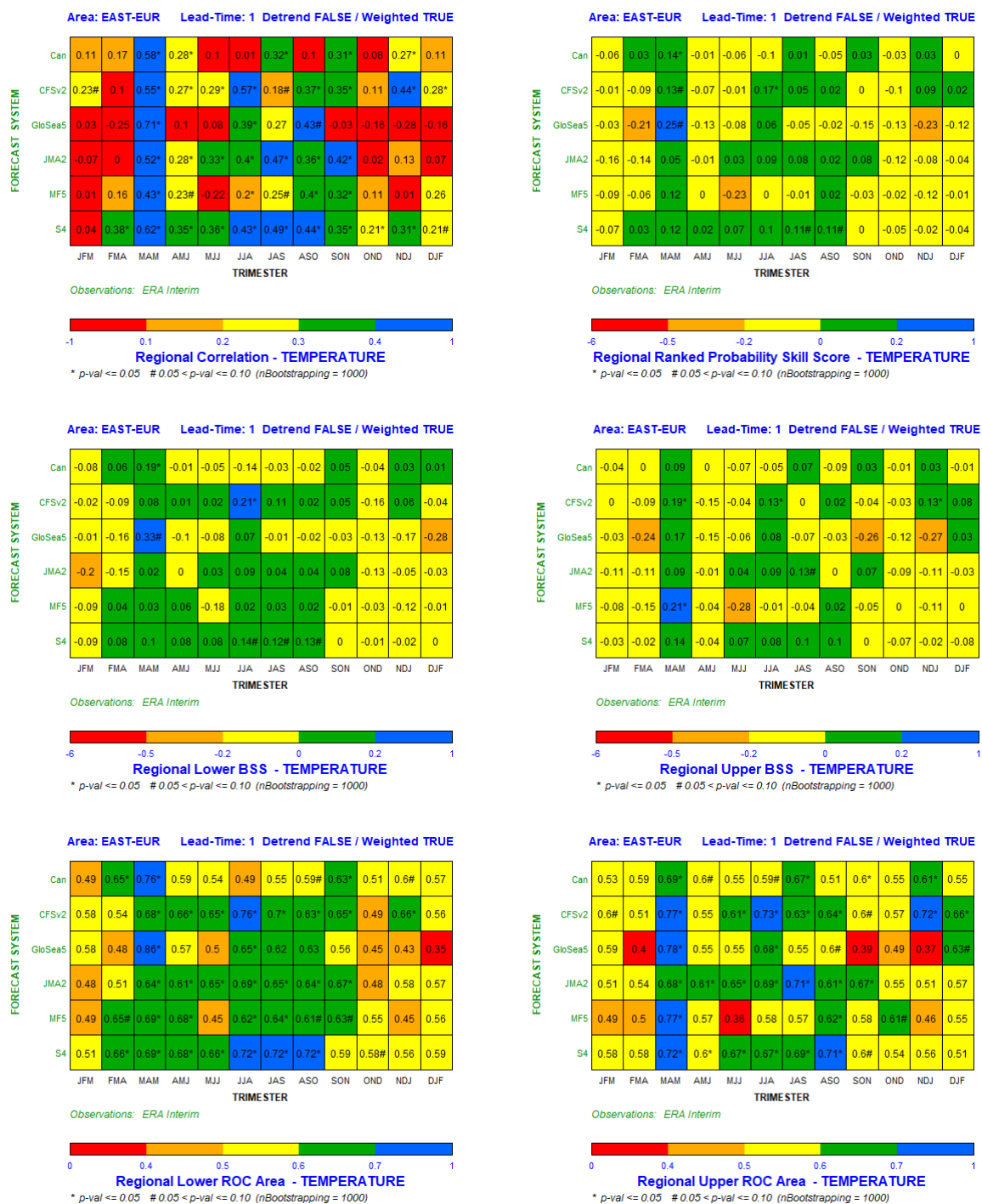


Table 36. The same as Table 32, but for EAST EUROPE domain

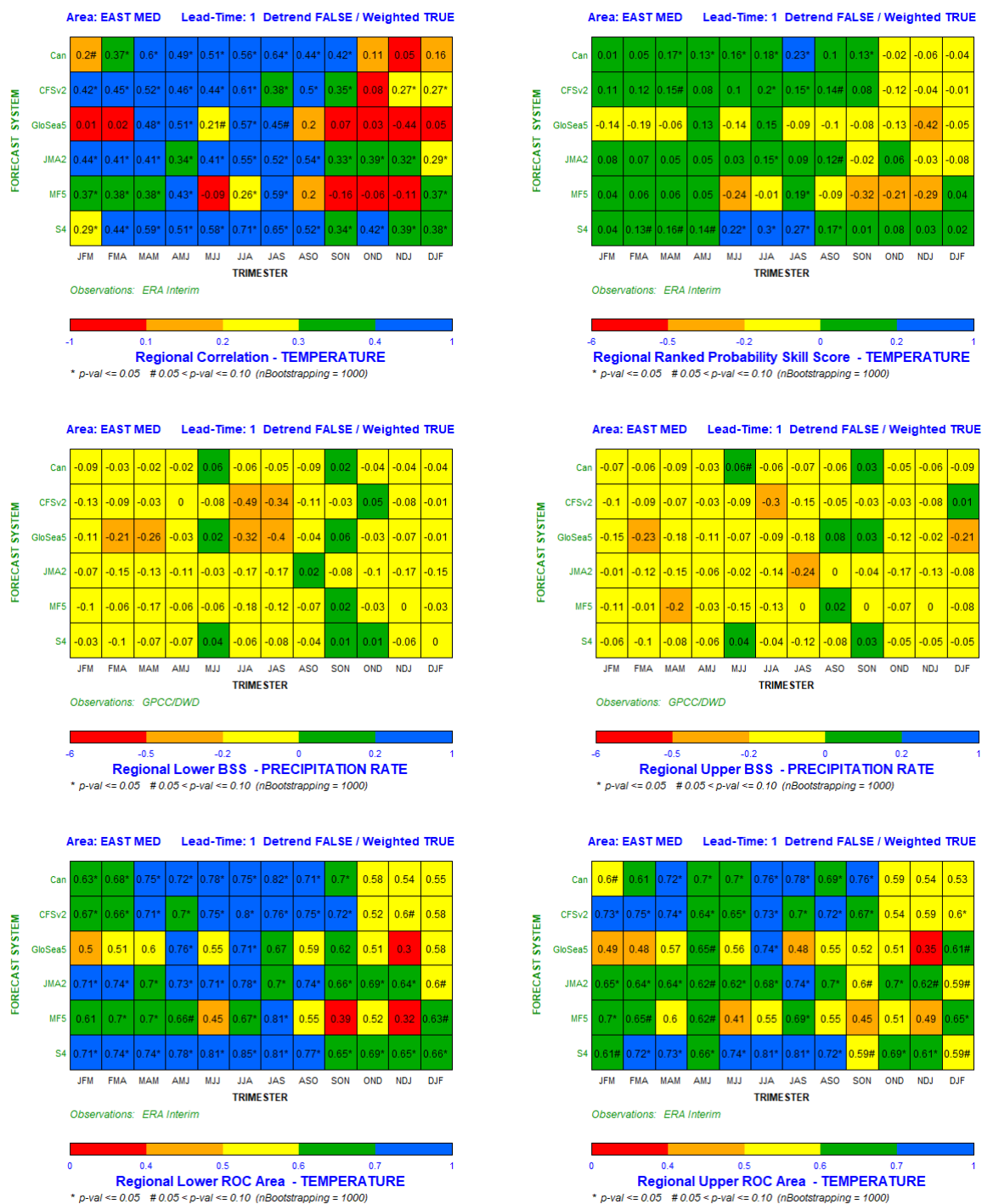


Table 37. The same as Table 32, but for EAST MEDITERRANEAN domain

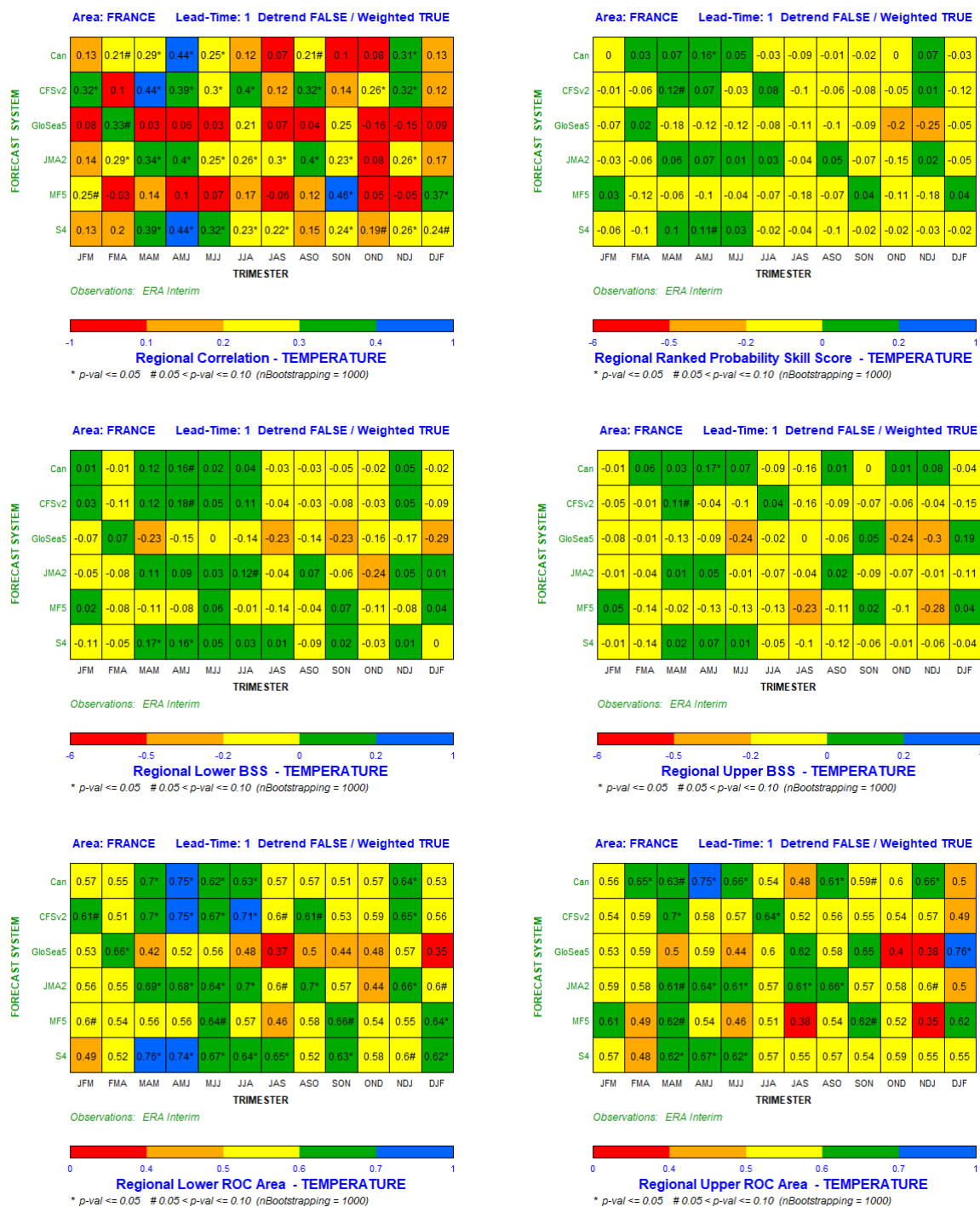


Table 38. The same as Table 32, but for FRANCE domain

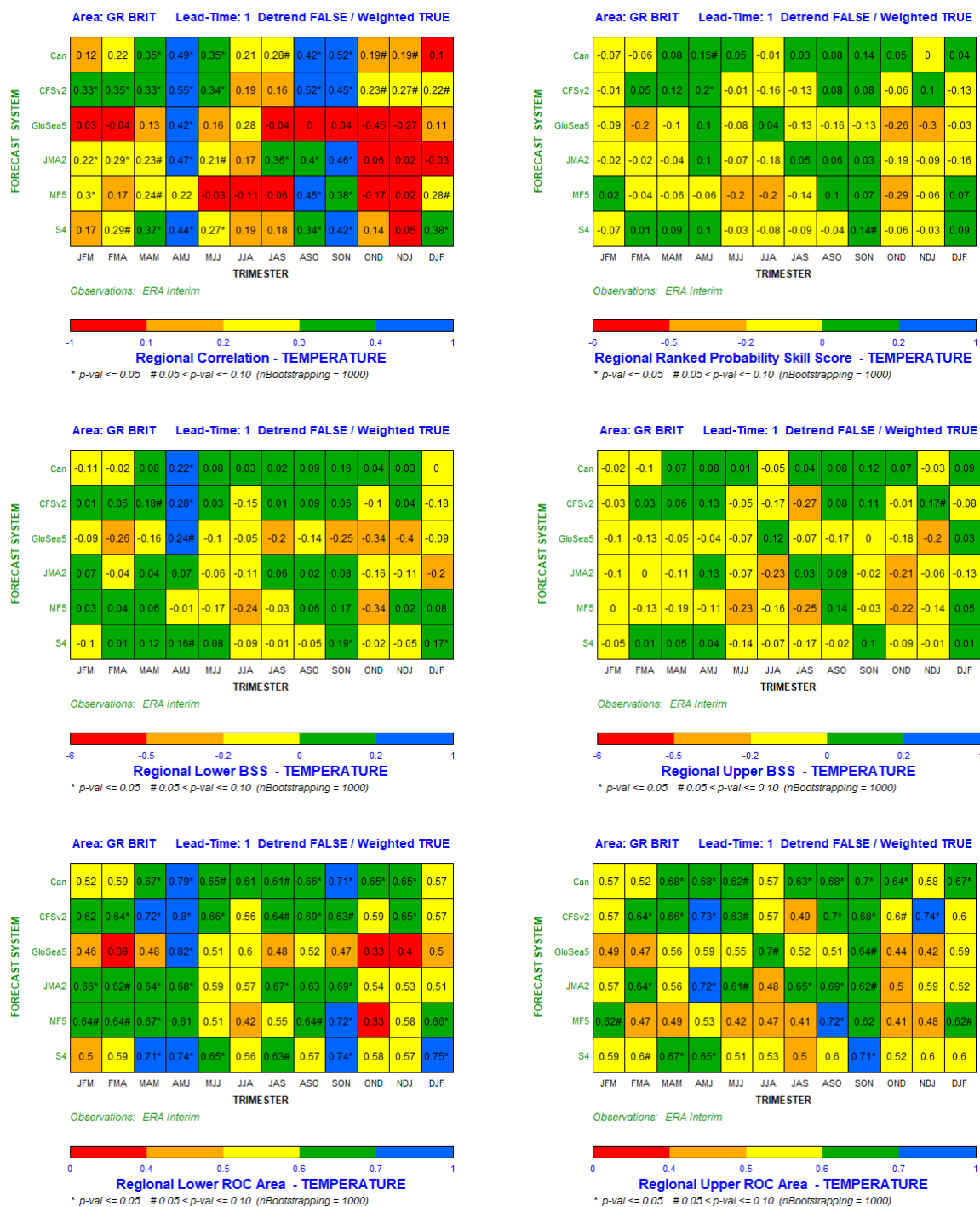


Table 39. The same as Table 32, but for GREAT BRITAIN domain

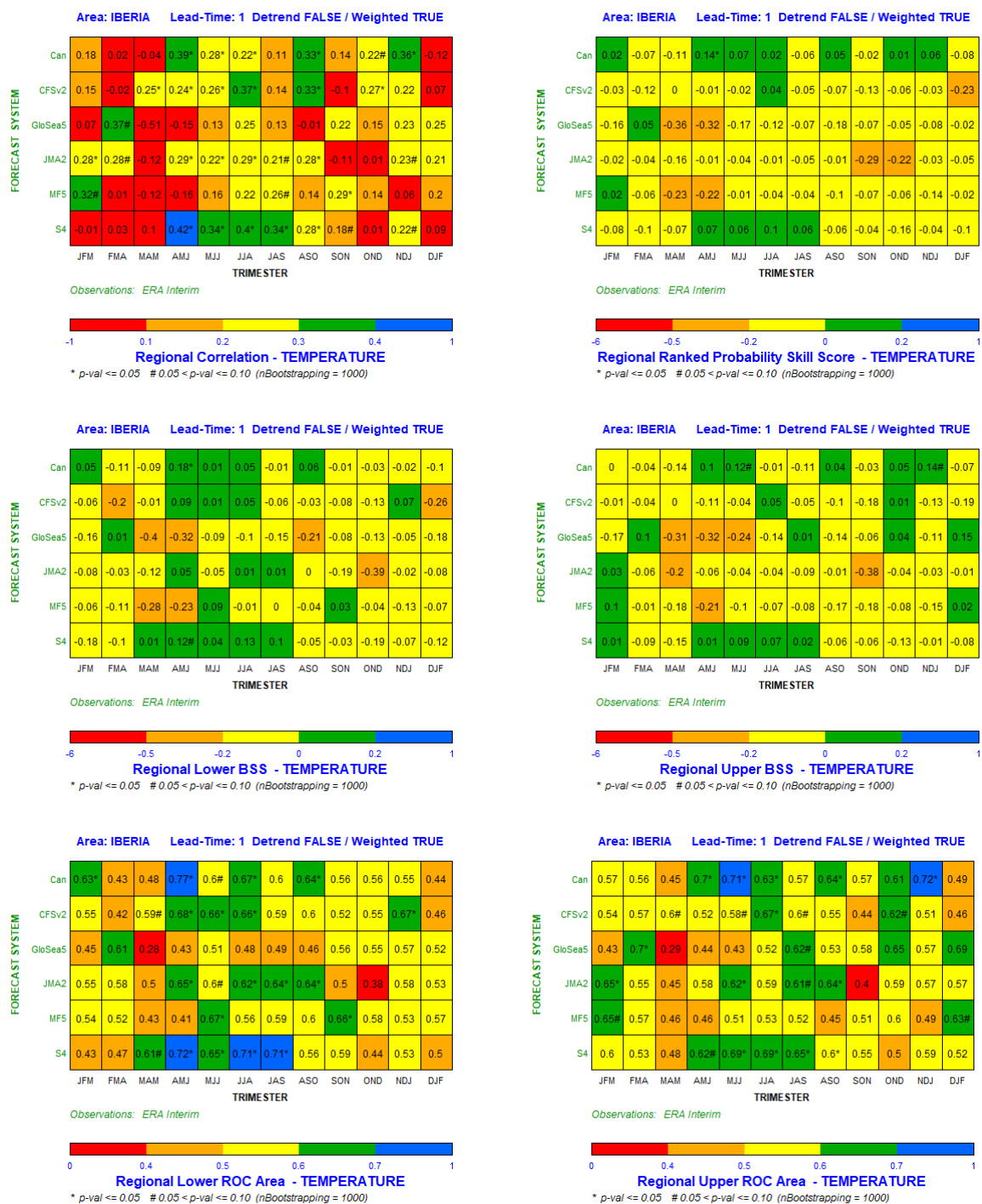


Table 40. The same as Table 32, but for IBERIA domain

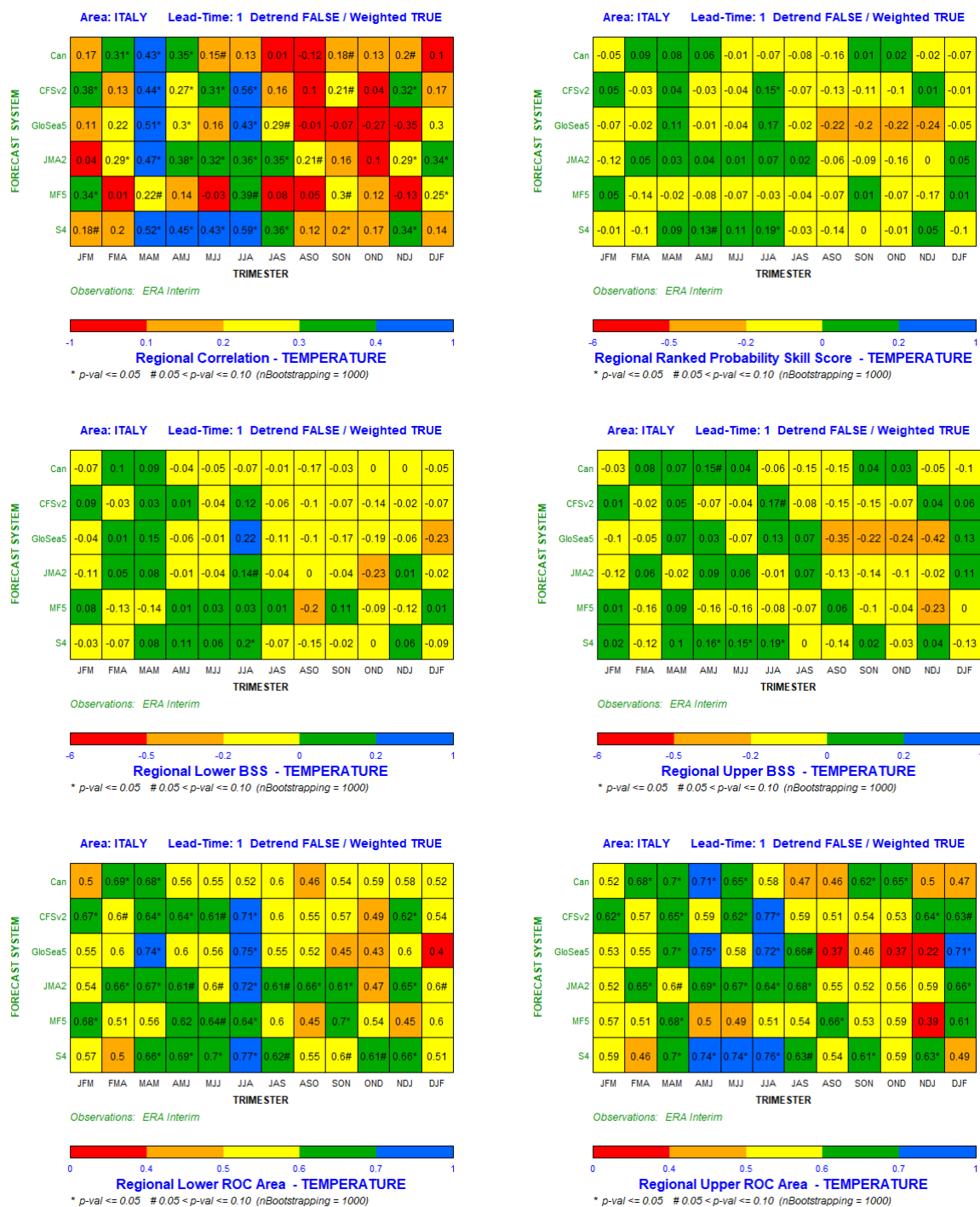


Table 41. The same as Table 32, but for ITALY domain

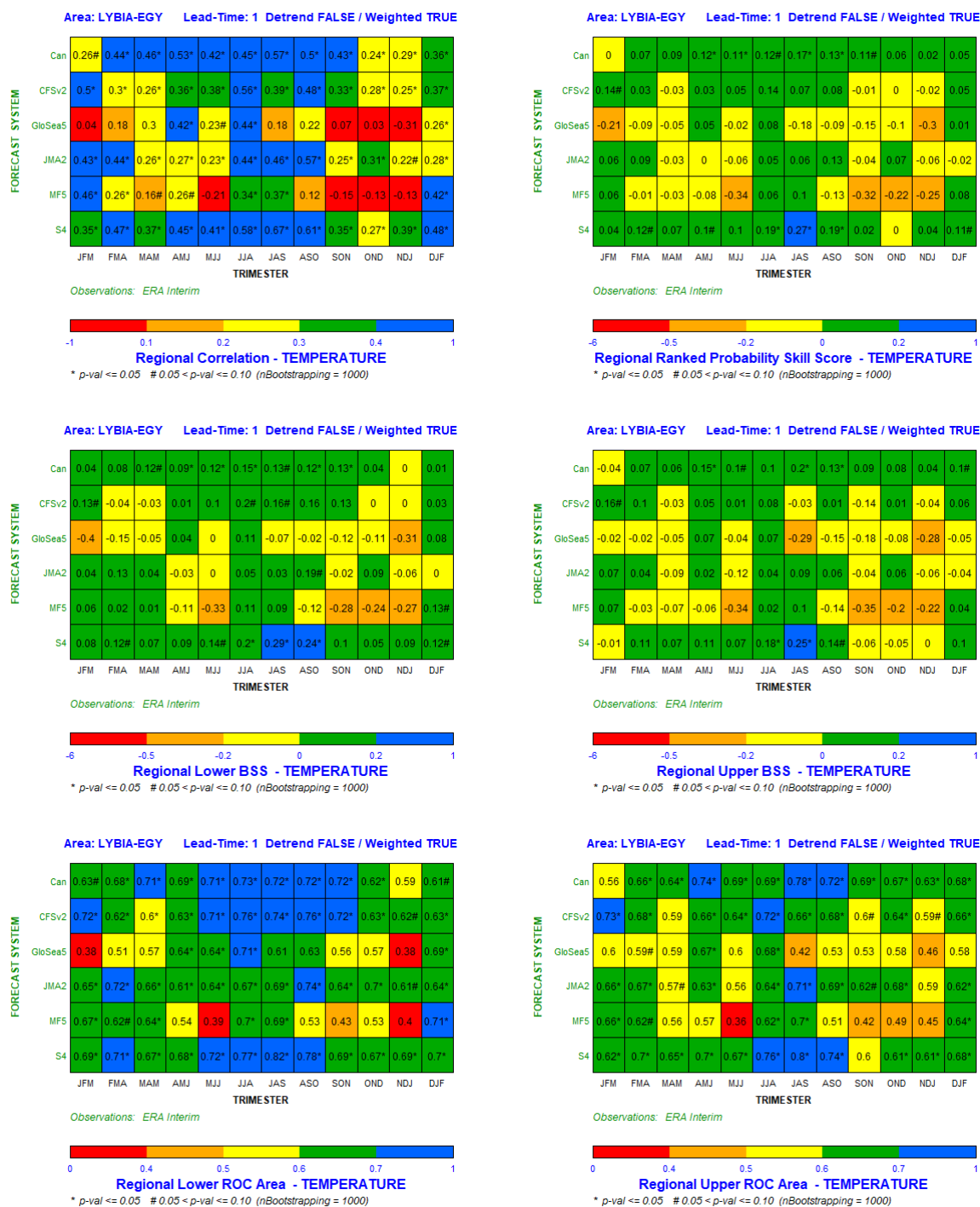


Table 42. The same as Table 32, but for LIBYA-EGYPT domain

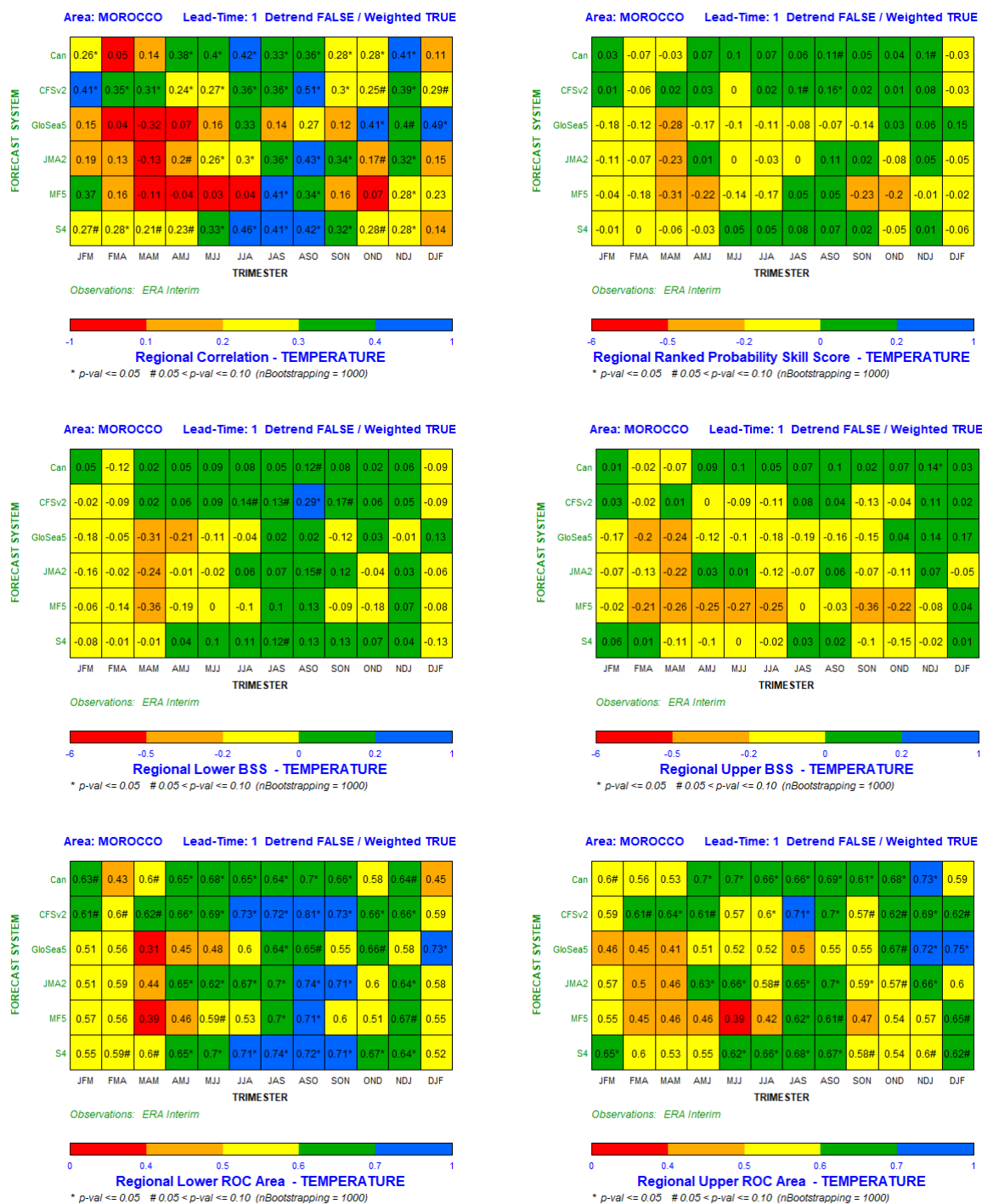


Table 43. The same as Table 32, but for MOROCCO domain

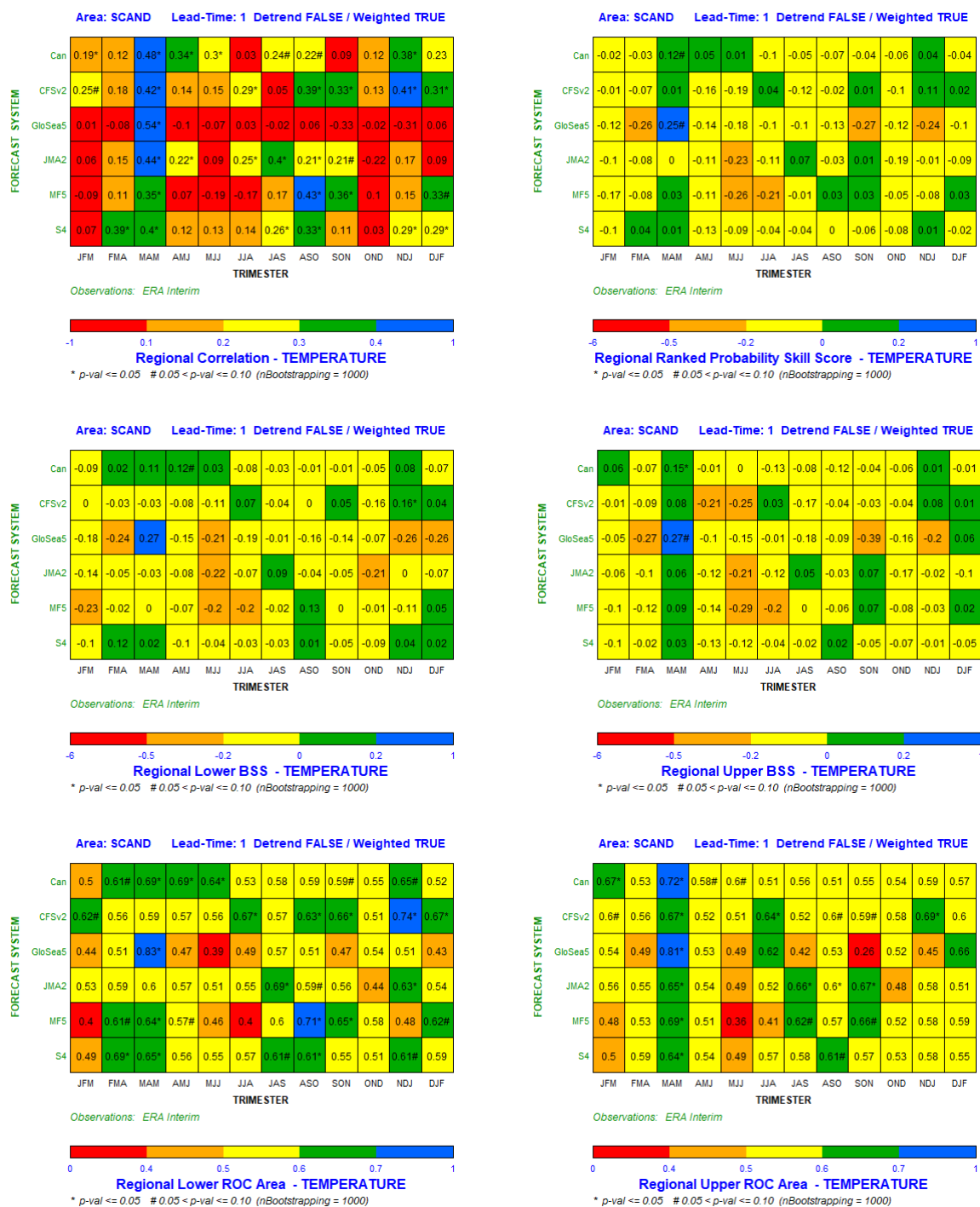


Table 44. The same as Table 32, but for SCANDINAVIA domain

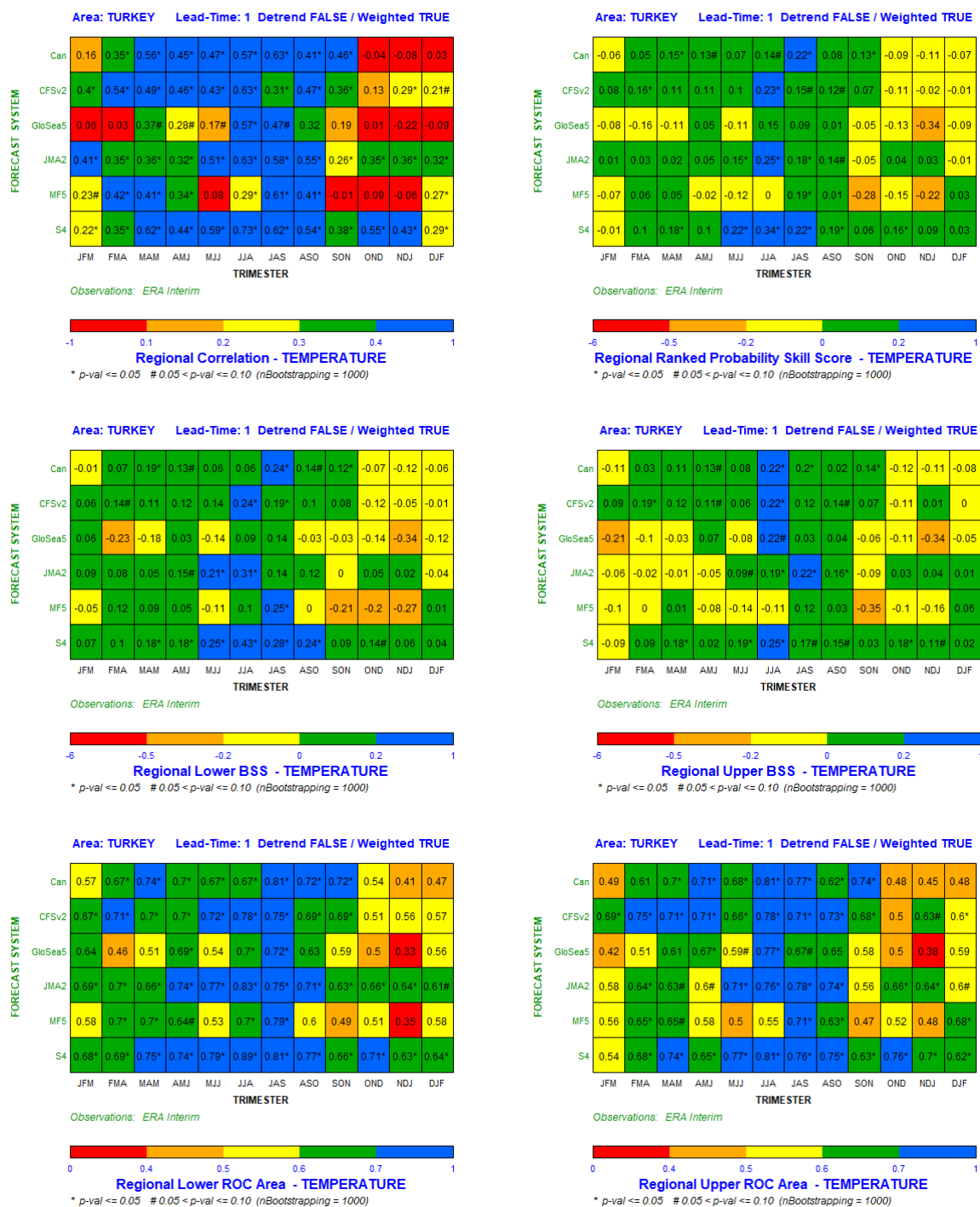


Table 45. The same as Table 32, but for TURKEY domain

4.2.b. Precipitation

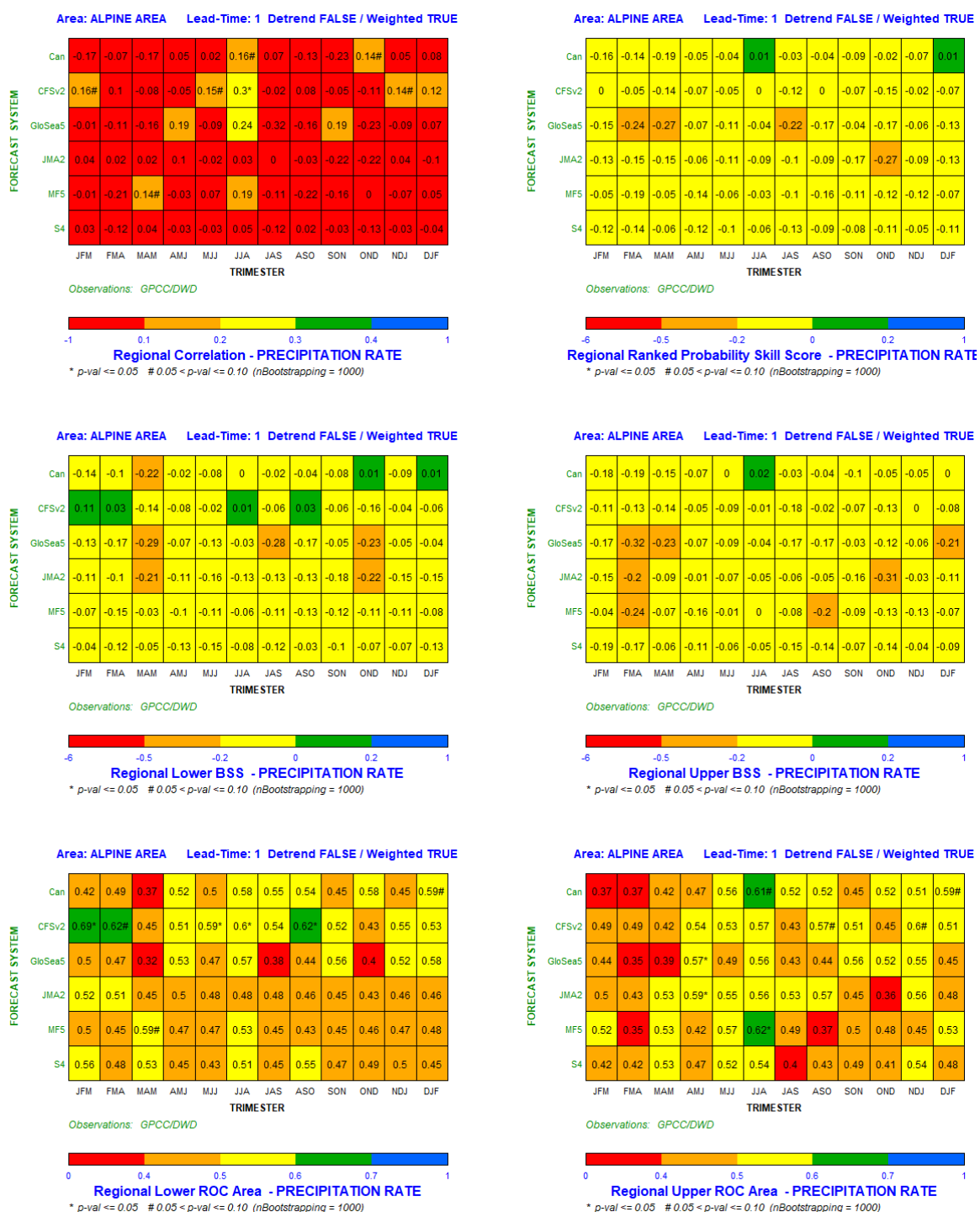


Table 46. Regional correlation coefficient, Ranked Probability Skill Score, Lower and Upper ROC Area and Brier Skill Score computed for precipitation anomalies (without removing trend), for 12 different three-month periods and lead-time 1 over ALPINE AREA domain (maximum available verification period for each system (see Table 1)). Three-month periods are shown in the X-axis and systems (S4, MF5, JMA2, GLOSEA5, CFSv2 and Can) are represented in the Y-axis (see text for description).

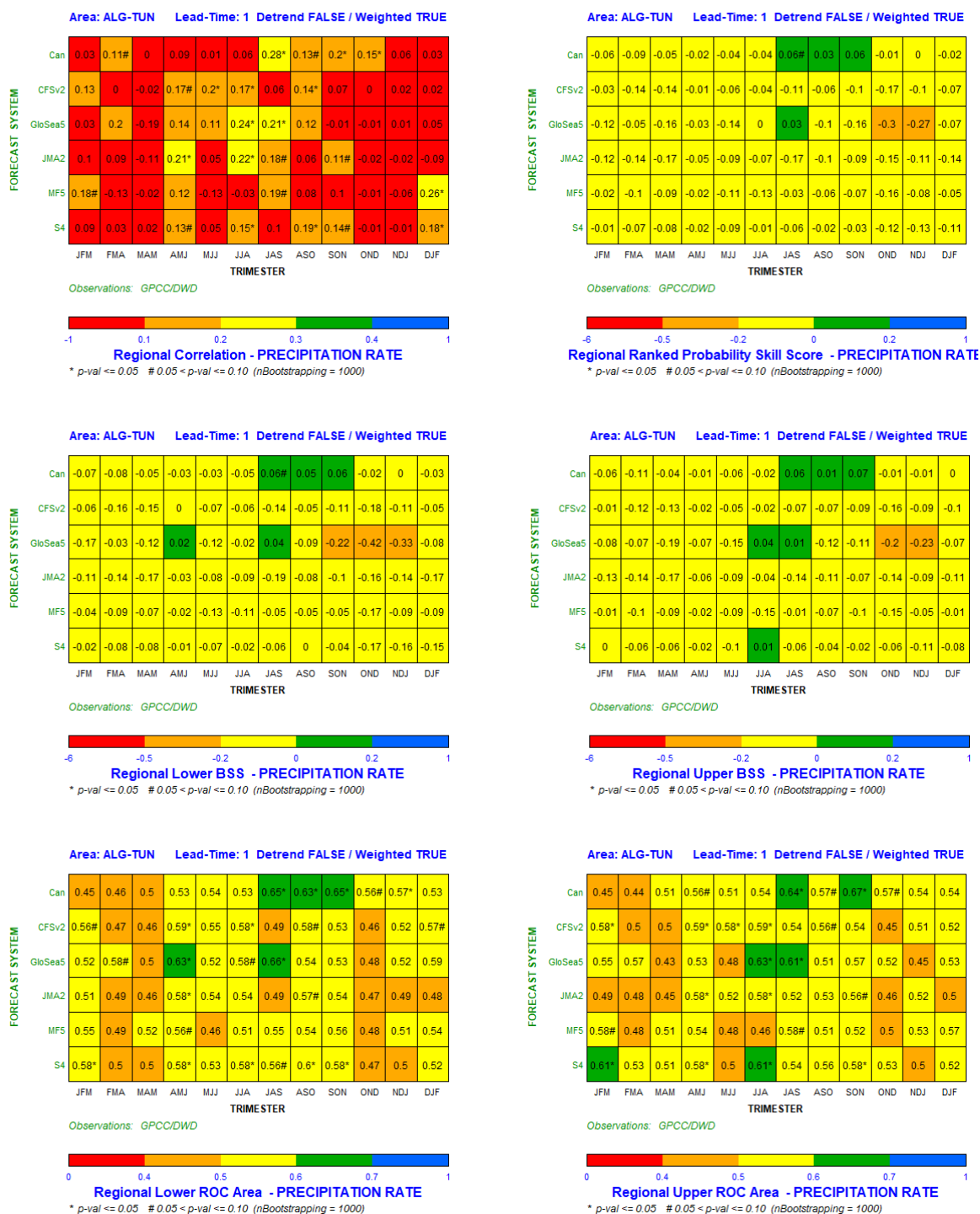


Table 47. The same as Table 46, but for ALGERIA/TUNISIA domain

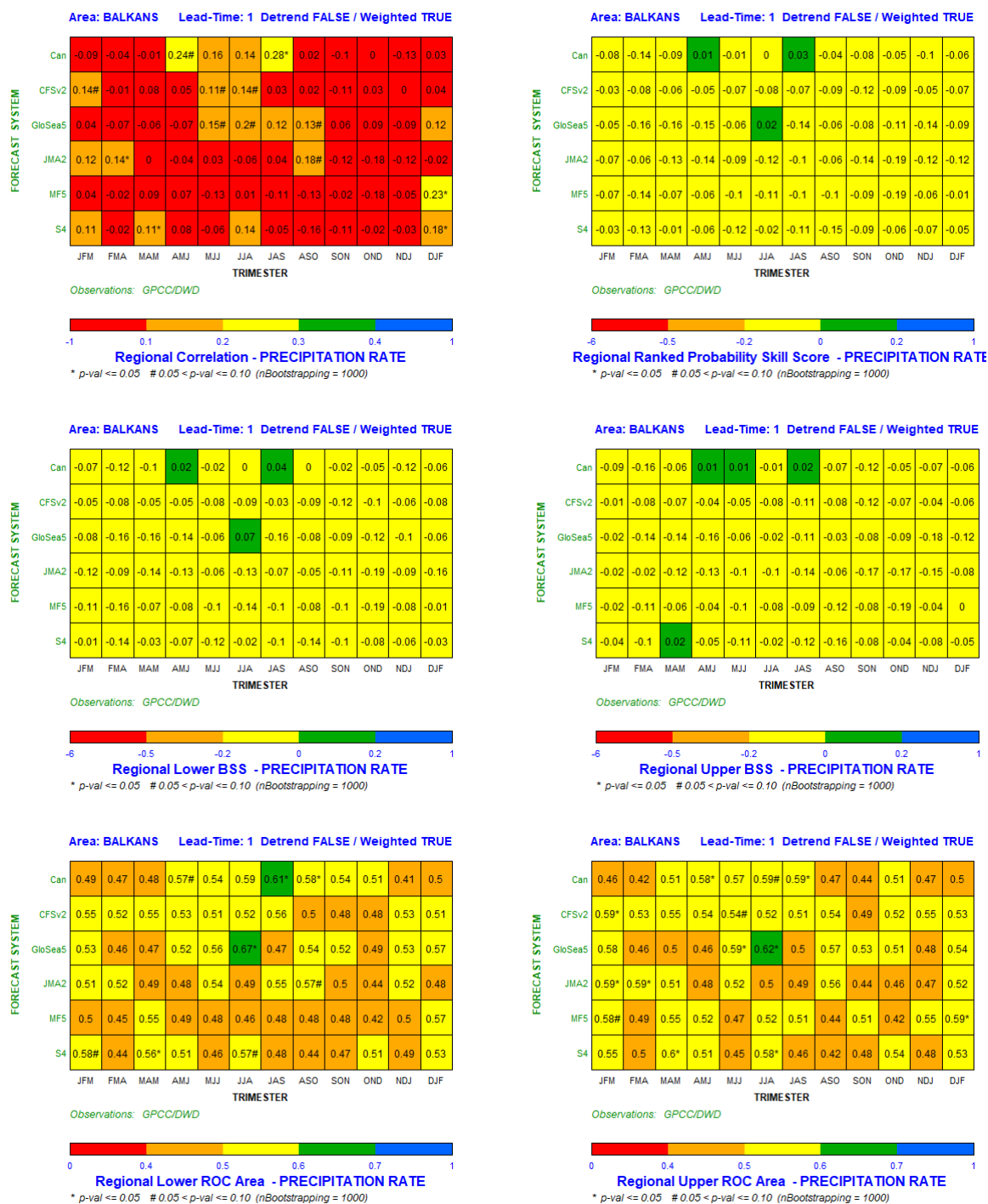


Table 48. The same as Table 46, but for BALKANS domain

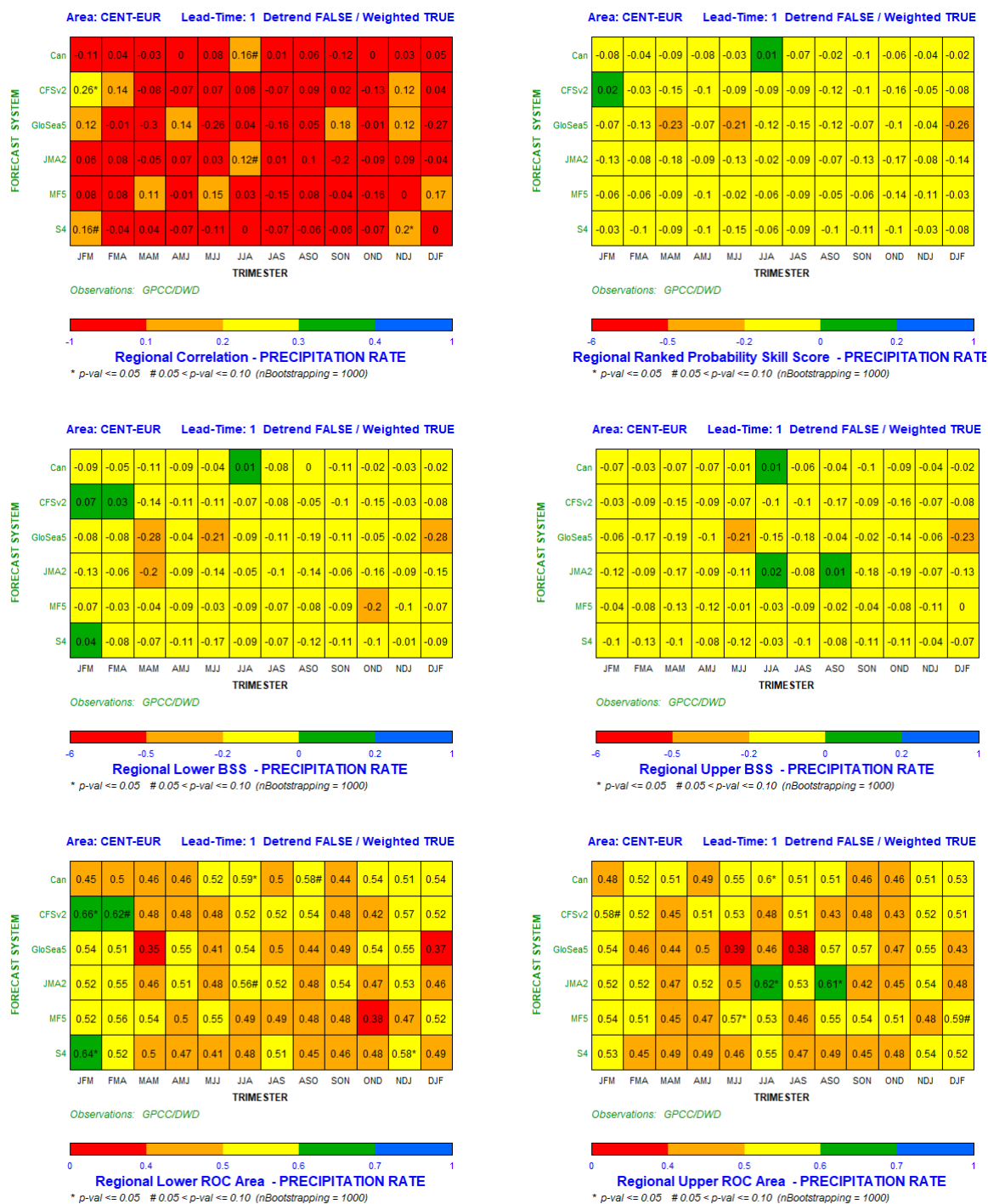


Table 49. The same as Table 46, but for CENTRE EUROPE domain

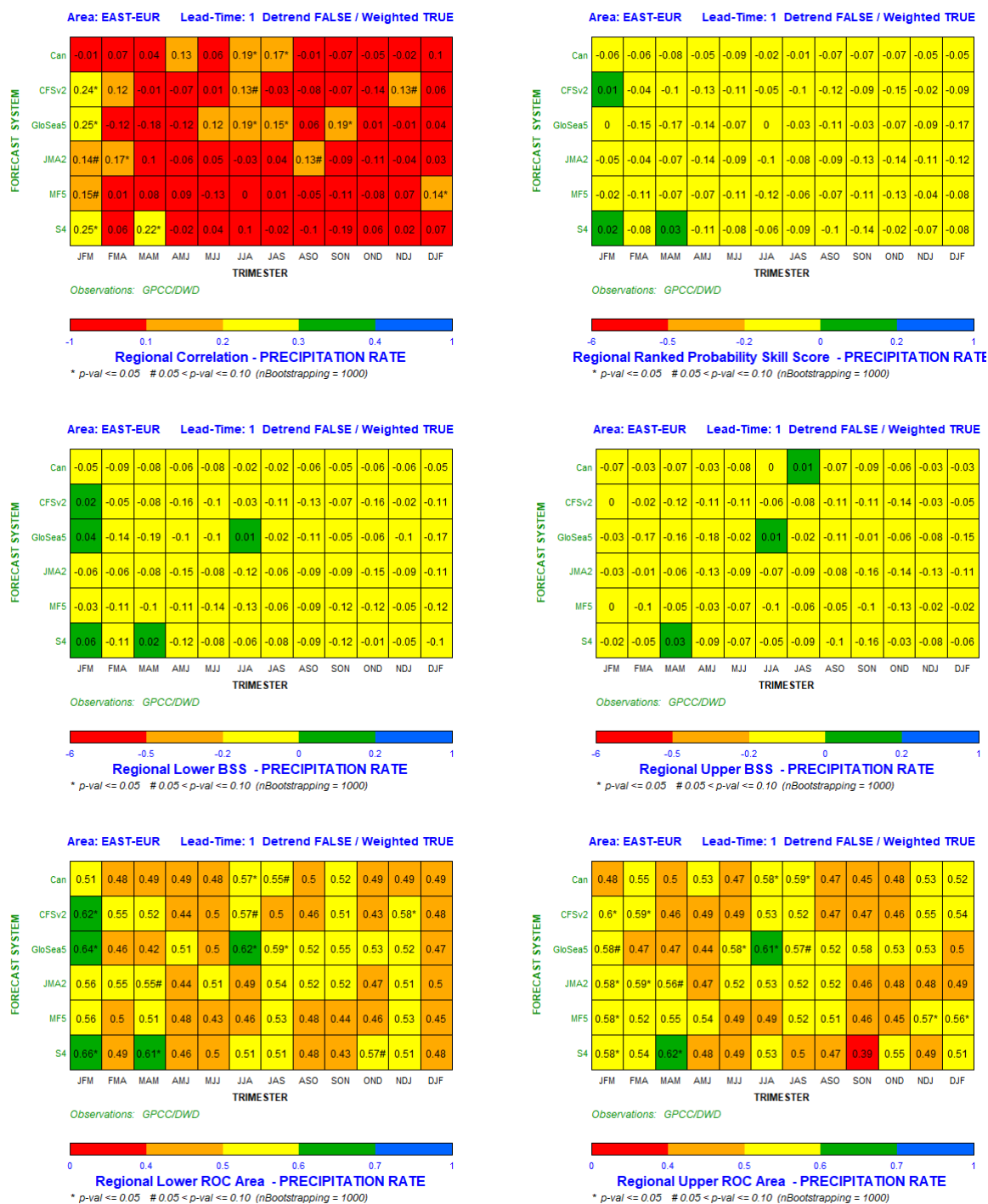


Table 50. The same as Table 46, but for EAST EUROPE domain

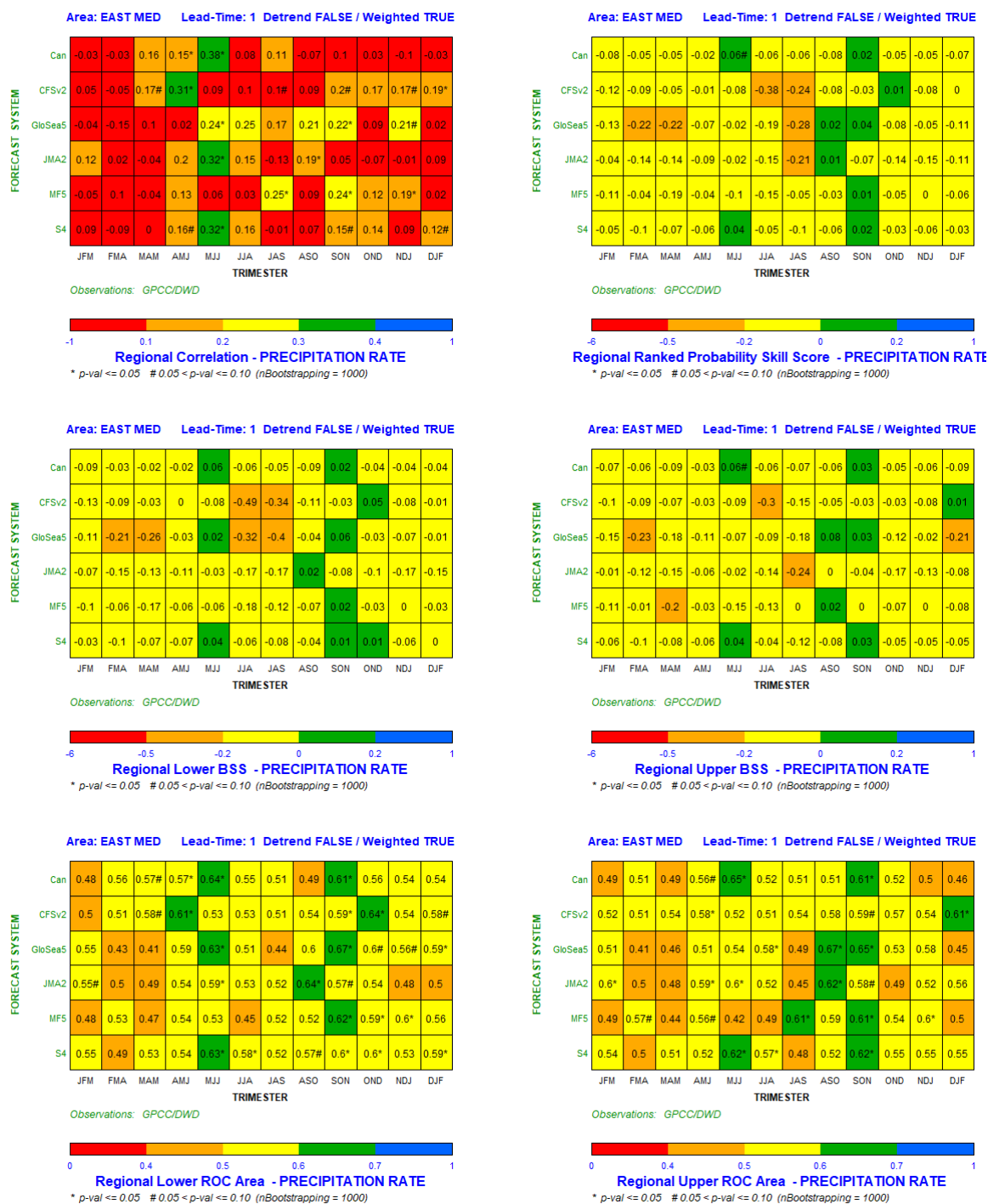


Table 51. The same as Table 46, but for EAST MEDITERRANEAN domain

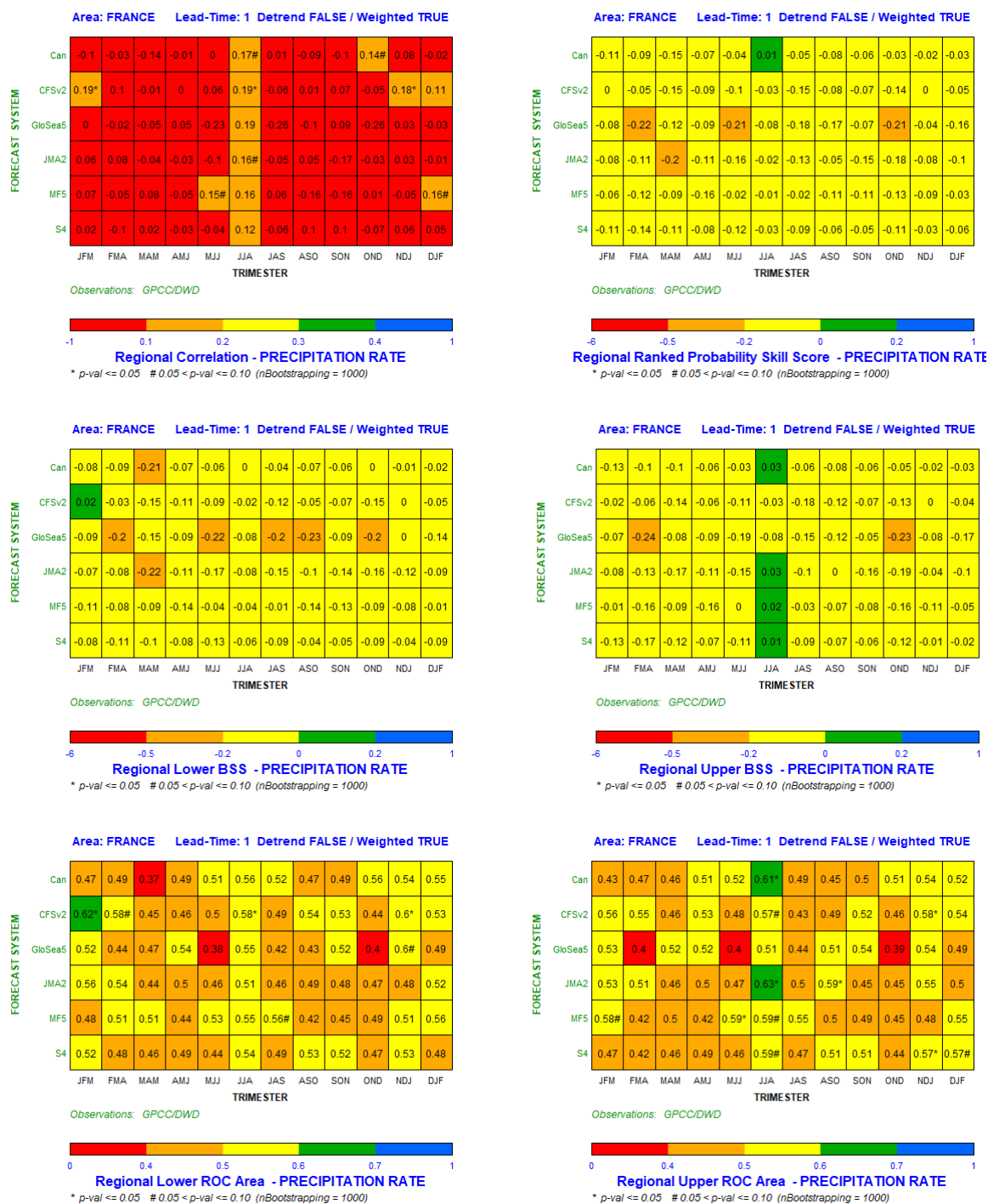


Table 52. The same as Table 46, but for FRANCE domain

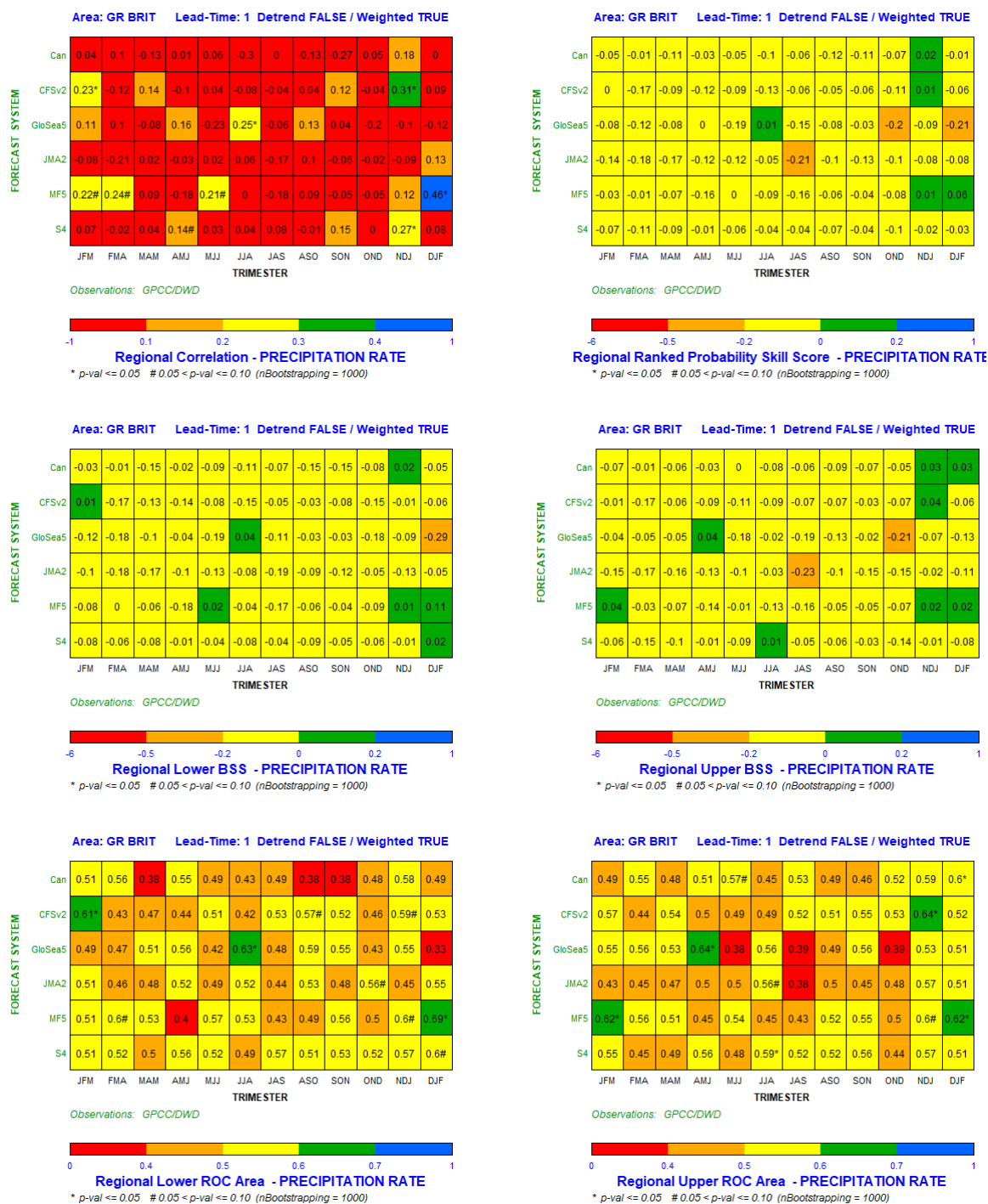


Table 53. The same as Table 46, but for GREAT BRITAIN domain

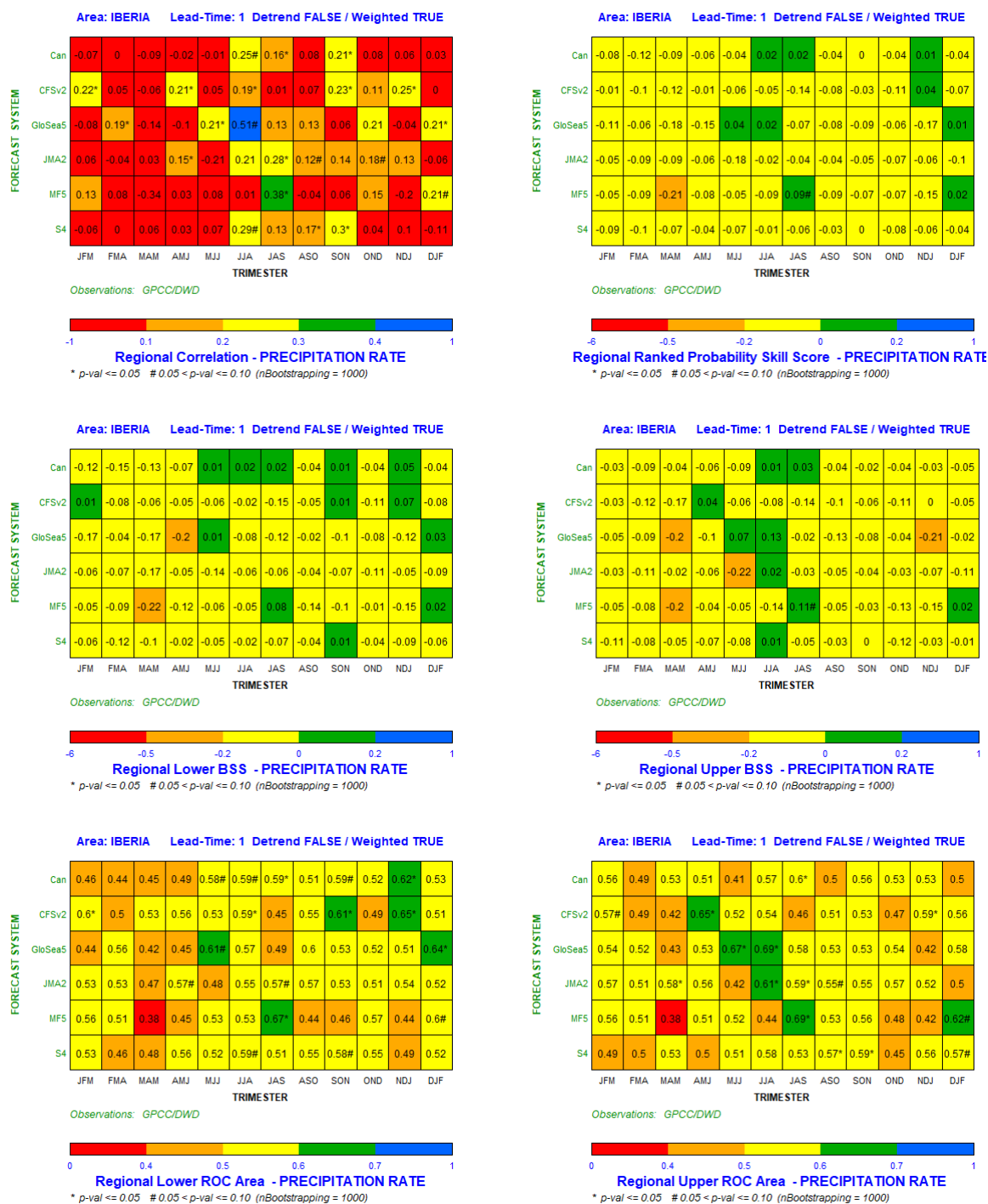


Table 54. The same as Table 46, but for IBERIA domain

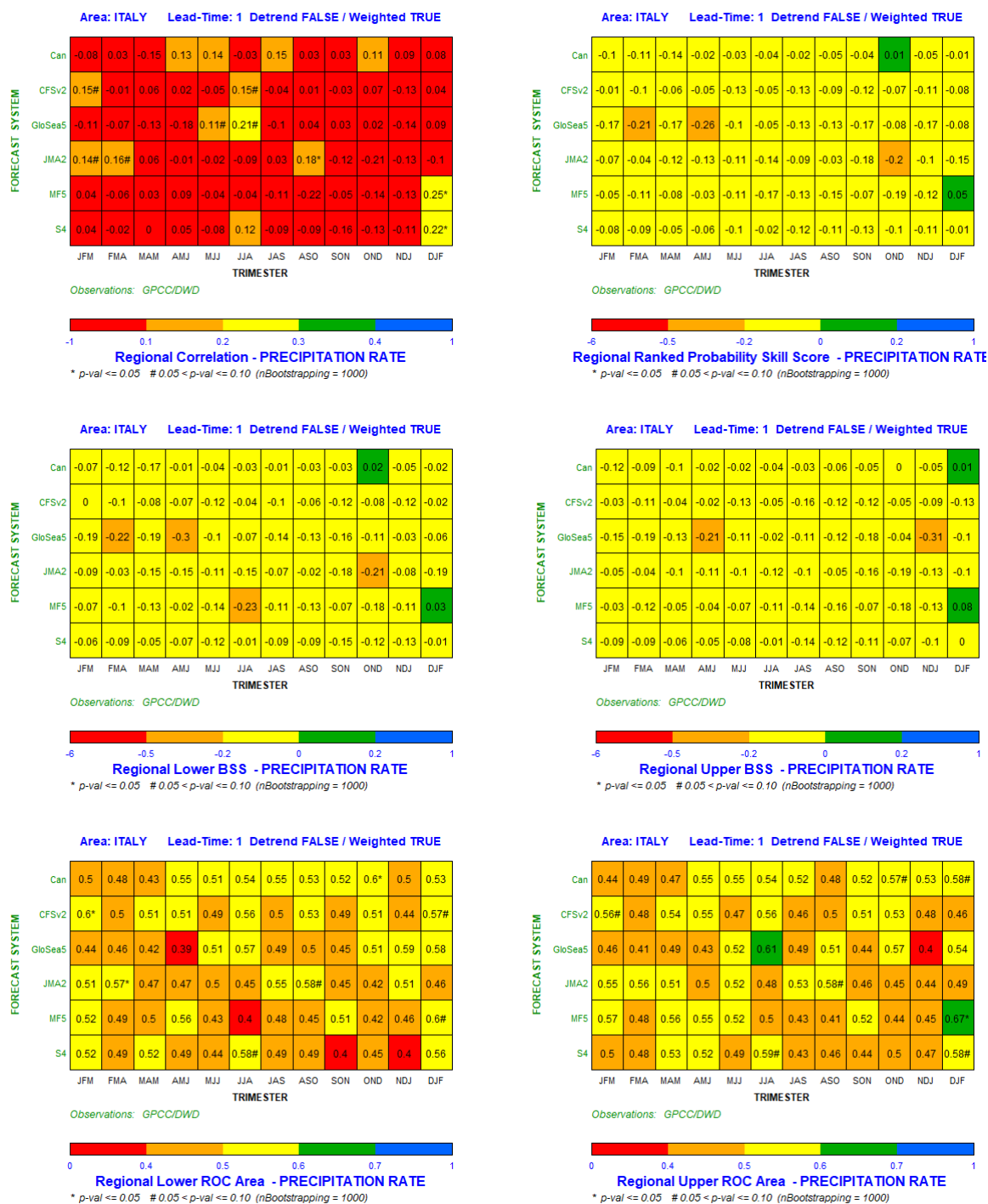


Table 55. The same as Table 46, but for ITALY domain

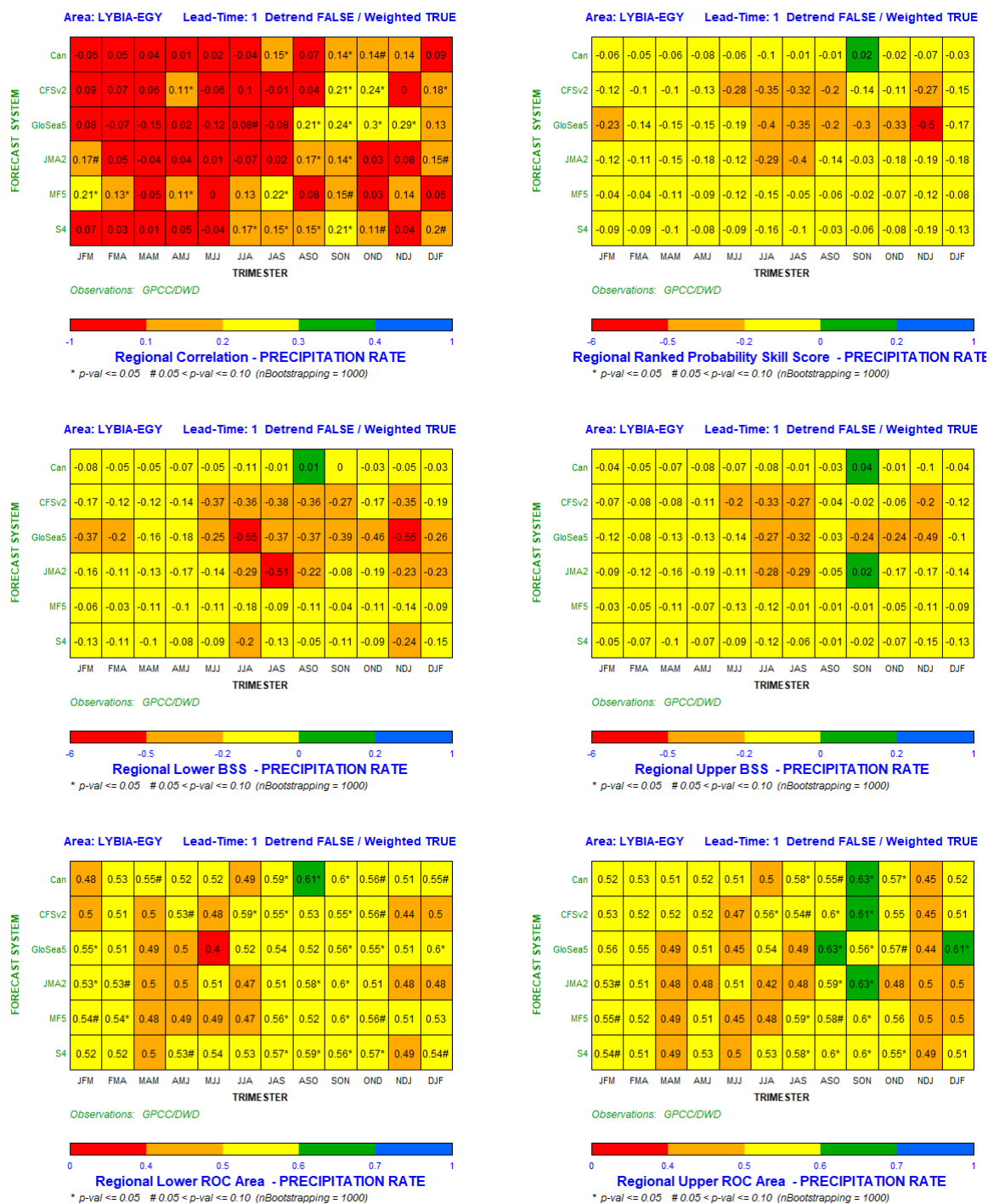


Table 56. The same as Table 46, but for LIBYA-EGYPT domain

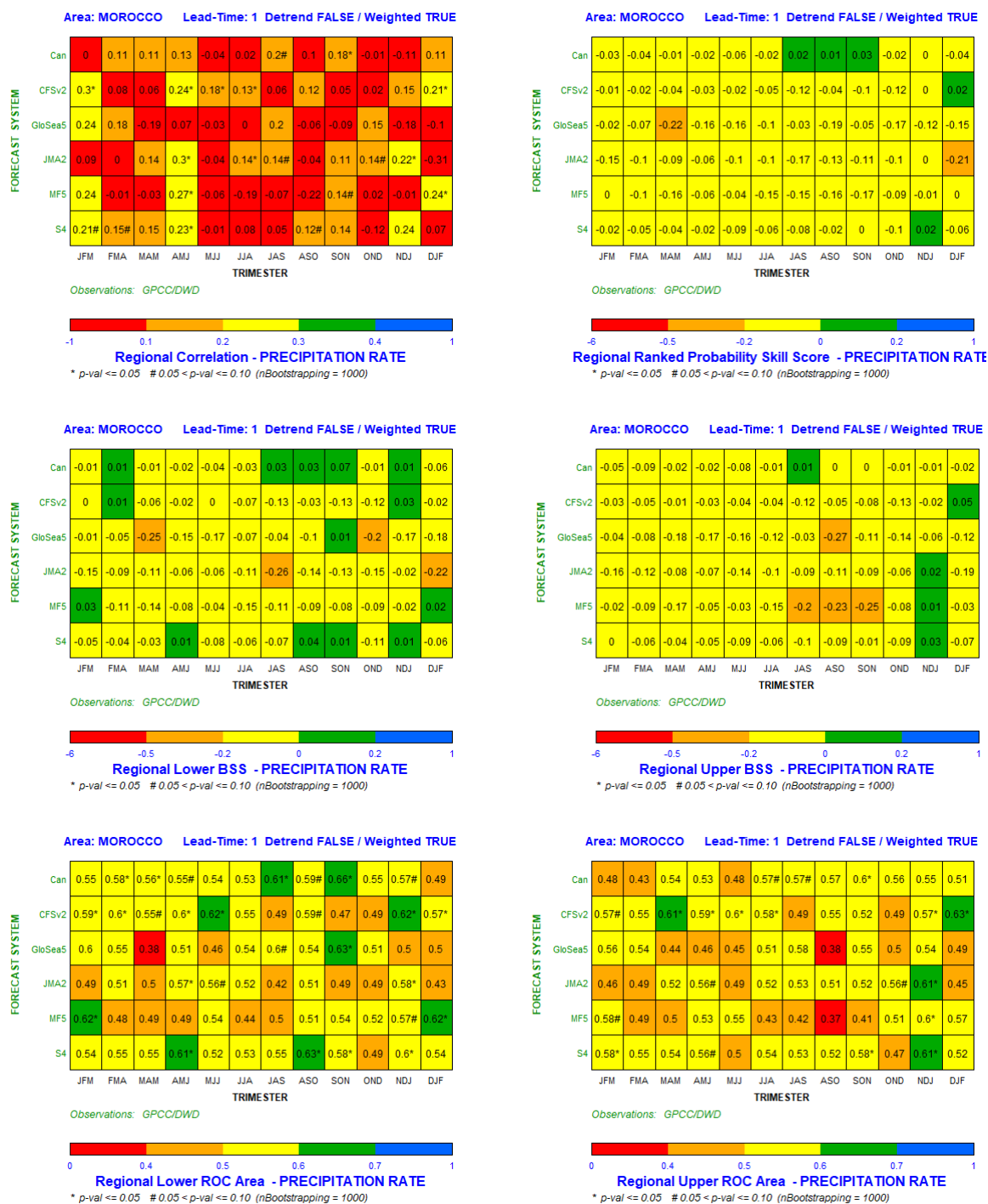


Table 57. The same as Table 46, but for MOROCCO domain

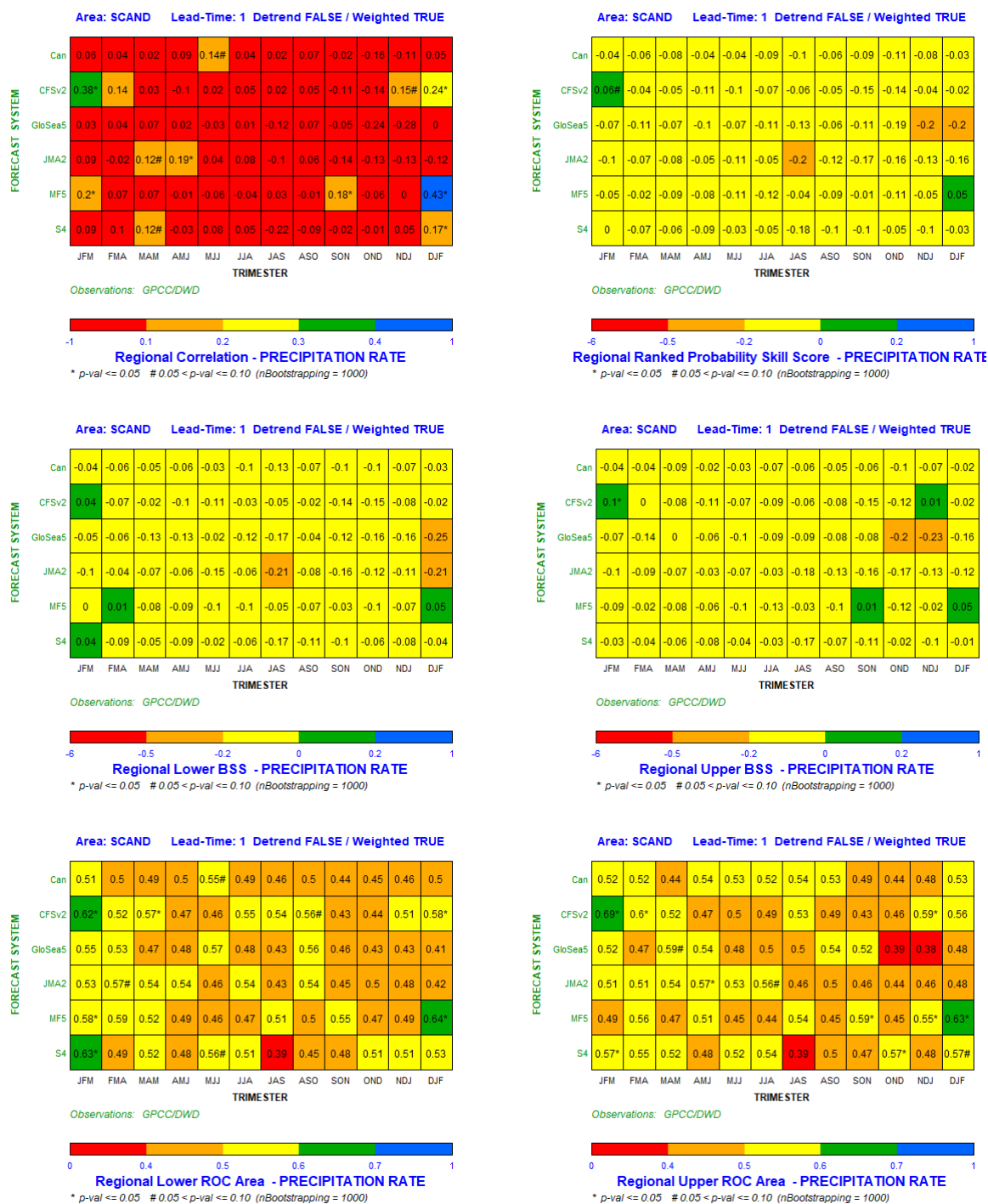


Table 58. The same as Table 46, but for SCANDINAVIA domain

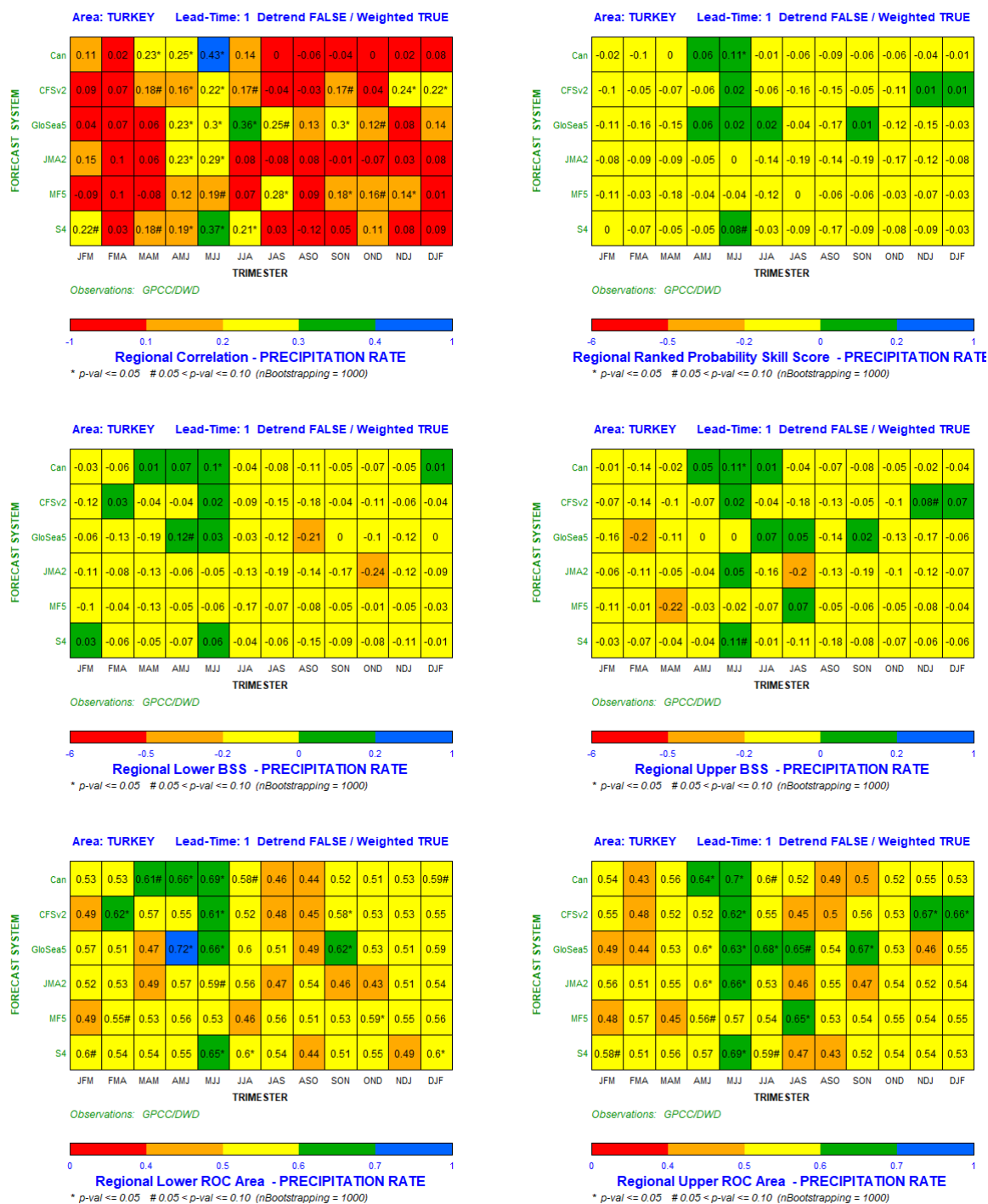


Table 59. The same as Table 46, but for TURKEY domain

5. SUMMARY AND CONCLUSIONS

Predictability of seasonal forecasts for surface variables such as temperature and precipitation is related to the low-frequency patterns. In spite of the steady advances, seasonal forecasting has to face many theoretical and practical challenges to become an everyday tool in the extratropical regions.

Predictability at seasonal time scale is limited in the extratropics and consequently dynamical seasonal forecasts skill is low over mid-latitude regions. Although verification results show, as expected, low skill at seasonal timescale, we can still draw the following general conclusions:

- Correlations and probabilistic scores are indeed better for temperature than for precipitation, as expected.
- Southern and Eastern domains show higher skill than Northern areas (e.g. Scandinavia vs East-Mediterranean domains).
- Relative consistency among models allows identification of some windows of opportunity for seasonal forecasts associated to certain seasons and variables; though in some cases are limited to certain models. Spring-Summer appears as a window of opportunity for temperature, possibly linked to the general trend associated to the climate system warming.
- Verification period has a strong influence on verification scores.
- Skill differs among scores, e.g., the ROC area skill score -providing an indication of a forecasts system discrimination ability- generally tends to show more skill (relative to climatology) than others exploring different aspects of forecasts.
- Precipitation scores tend to show big differences among domains. Best results generally correspond to the Eastern domain. Autumn shows remarkable good scores for all models over East-Mediterranean domain.
- When using the largest available verification periods, temperature tend to show for all models better scores than climatology in spring and summer seasons, specifically as regards scores measuring resolution. Scores tend to be higher when trend is not removed. There is also a clear difference between Eastern and Southern Mediterranean (Algeria-Tunisia, Libya-Egypt, East-Mediterranean and Turkey) and the rest of domains.
- Removal of temperature trend seems not to have significant impact on the verification scores for the short common verification period (1997-2009). However, for the maximum available period for each model, values decrease significantly for certain seasons although most models show ROC area scores higher than 0.5. In particular, this effect is lower for the MF5 model with the shorter hindcast period, whereas no effect appears for GloSea5 model.

Update of verification scores for the latest seasonal forecasting system versions is always a need for supporting operational activities. Verification scores here described will be complemented in the future with the corresponding scores for the coming versions. This comparison exercise of models and versions skill may also shed light on sources of predictability linked to certain seasons or regions.

REFERENCES

- Coelho, C.A.S., Stephenson, D.B., Balmaseda, M., Doblas-Reyes, F.J and van Oldenborgh, G. J. (2006). “Towards an integrated seasonal forecasting system for South America”. *Journal of Climate*, Volume 19, Issue 15, 3704-3721.
- Curry, J.A. and Webster, P.J. (2011). “Climate Science and the Uncertainty Monster”. American Meteorological Society. (<https://doi.org/10.1175/2011BAMS3139.1>)
- Doblas-Reyes, F.J. (2010). “Seasonal prediction over Europe. Proceedings of the ECMWF”. Seminar on Predictability in the European and Atlantic regions.
- Gneiting, T. and Raftery, A.E. (2005). “Atmospheric science. Weather forecasting with ensemble methods”. *Science*, Volume 310, Issue 5746, 248-249.
- Jolliffe, I.T. and Stephenson, D.B. (2003). “Forecast Verification. A Practitioner’s Guide in Atmospheric Science”. ISBN 0-471-49759-2.
- Kim, H.M, Webster, P.J. and Curry, J.A (2012). “Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter”. *Climate Dynamics*, Volume 39, Issue 12, 2957–2973.
- Kirtman, B. and Pirani, A. (2008). “WCRP Position Paper on Seasonal Prediction”. WCRP Informal Report No. 3/2008, ICPO Publication No. 127.
- Knutti, R. (2010). “Challenges in Combining Projections from Multiple Climate Models”. American Meteorological Society (<https://doi.org/10.1175/2009JCLI3361.1>)
- Molteni, F. et al. (2011). “The new ECMWF seasonal forecast system (System 4)”. ECMWF Technical Memoranda, 656.
- Murphy, A.H. and Winkler, R.L. (1984). “Probability Forecasting in Meteorology”. *Journal of the American Statistical Association* Volume 79, Issue 387, 489-500.
- Palmer, T.N. (2000). “Predicting uncertainty in forecasts of weather and climate”. *Reports on Progress in Physics*, Volume 63, Issue 2.
- Sanchez-Garcia, E., Voces-Aboy, J., Rodríguez-Camino, E. (2014). “Calibration and combination of seasonal forecast over Southern Europe”. Publicación AEMET. (http://www.aemet.es/documentos/es/conocermas/recursos_en_linea/publicaciones_y_estudios/publicaciones/CCPE_english/CCPE_SouthEurope_english.pdf)

- Slingo, J. and Palmer, T. (2011). “Uncertainty in weather and climate prediction”. Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences, Volume 369, Issue 1956, 4751-4767.
- Troccoli, A., Harrison, M., Anderson, D.L.T. and Mason, S.J. (2008) “Seasonal climate: Forecasting and Managing Risk”. Nato Science Series: IV, Earth and Environmental Sciences, ISSN: 1568-1238.
- Weisheimer, A. et al. (2009). “ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs”. Geophysical Research Letters, Volume 36, Issue 21 (<https://doi.org/10.1029/2009GL040896>).
- Weller, R.A., Anderson, J.L., Arribas, A., Dickinson, R.E., Goddard L., Kalnay, E., Kirtman, B., Koster, R.D., Richman, M.B., Saravanan, R., Waliser, D. and Wang, B. (Committee on Assessment of Intraseasonal to Interannual Climate Prediction and Predictability) (2010). “Assessment of Intraseasonal to Interannual Climate Prediction and Predictability”. National Research Council. National Academies Press. ISBN: 0-309-15184-8.
- Wilks, D. S. (2006) “Statistical methods in the atmospheric sciences” ISBN 13:978-0-12-751966-1.
- WMO. “Standardised Verification System for Long-Range Forecasts” New Attachment II-8 to the Manual on the GDPFS, Volume I, Issue 485.

ANNEX I

RESULTS FOR TEMPERATURE WHEN TREND IS REMOVED.

1. VERIFICATION PERIOD: 1997-2009

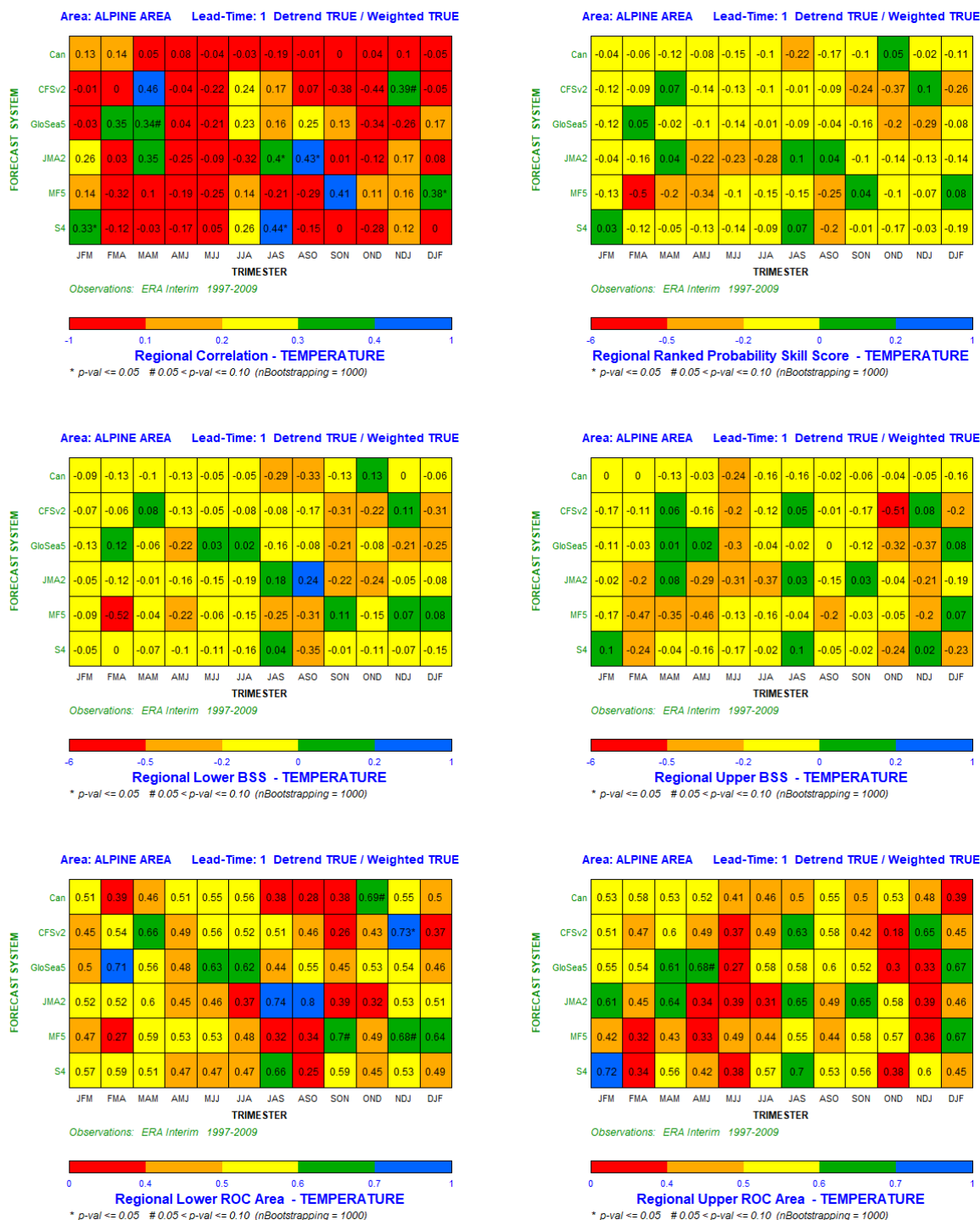


Table 60. Regional correlation coefficient, Ranked Probability Skill Score, Lower and Upper ROC Area and Brier Skill Score computed for temperature anomalies (removing trend), for 12 different three-month periods and lead-time 1 over ALPINE AREA domain (common verification period for all systems: 1997-2009). Three-month periods are shown in the X-axis and systems (S4, MF5, JMA2, GLOSEA5, CFSv2 and Can) are represented in the Y-axis (see text for description).

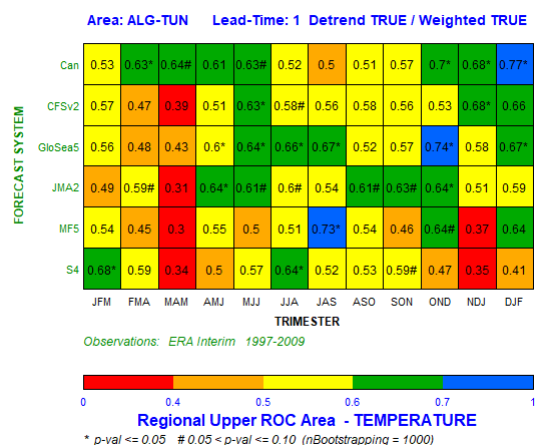
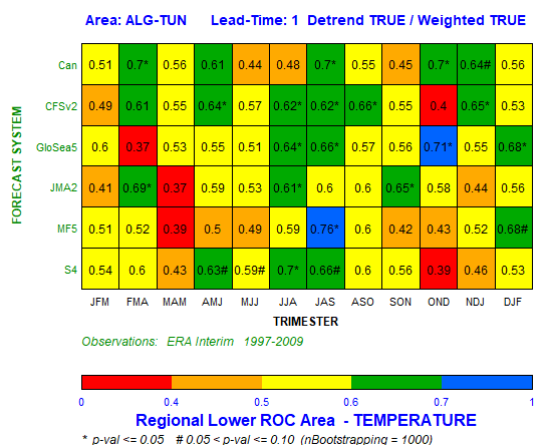
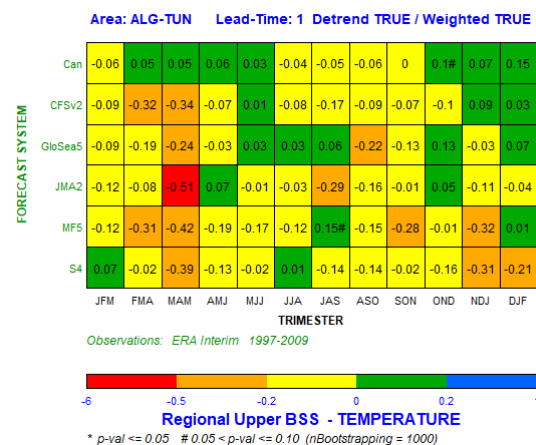
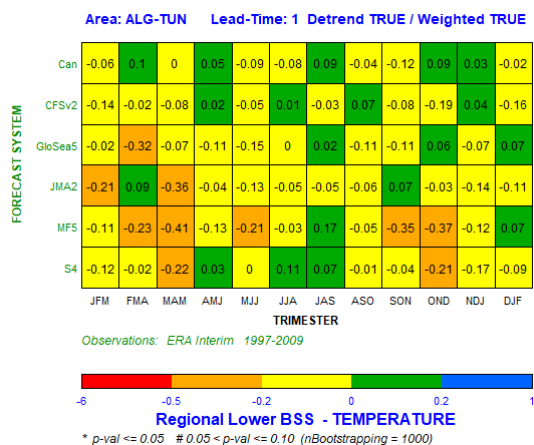
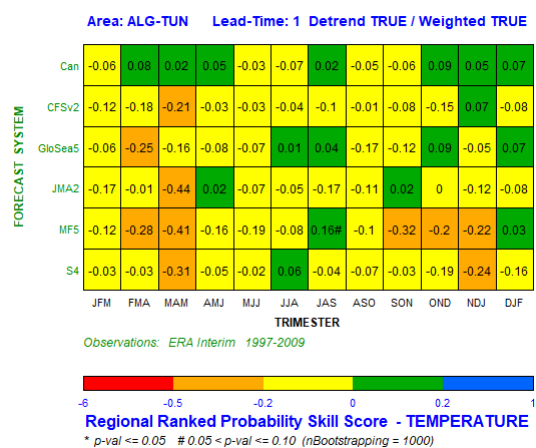
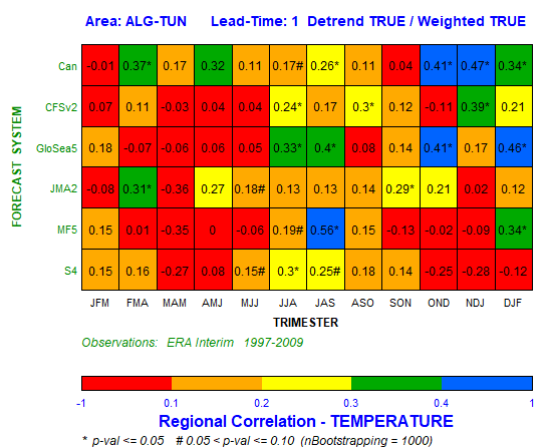


Table 61. The same as Table 60, but for ALGERIA/TUNISIA domain

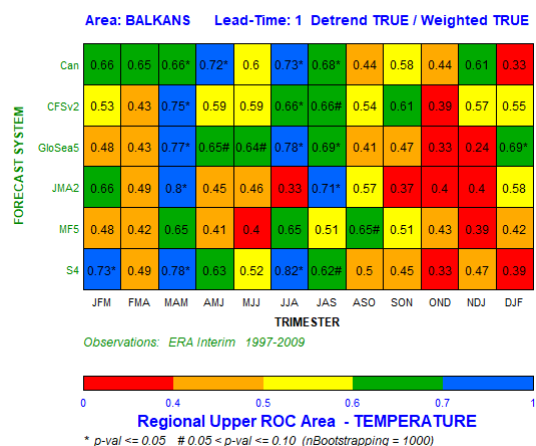
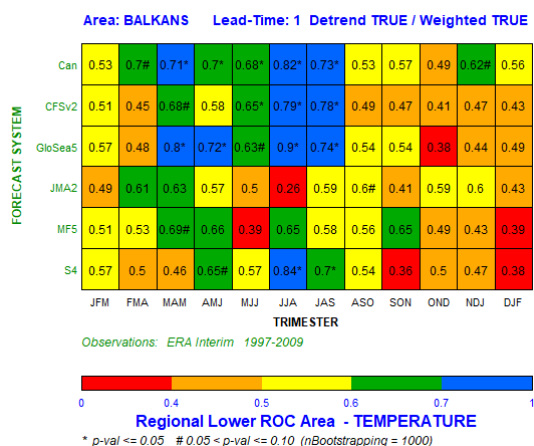
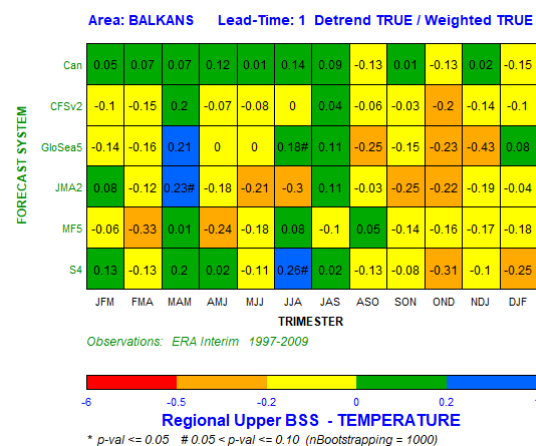
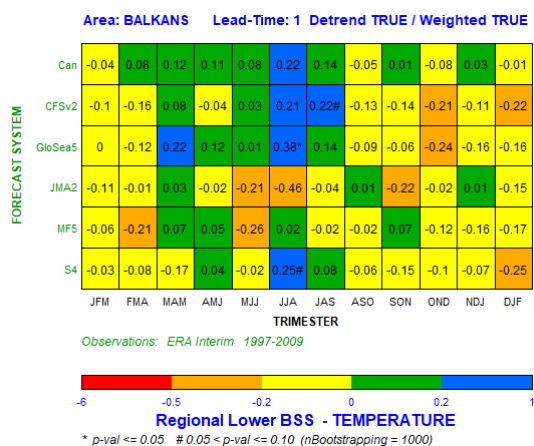
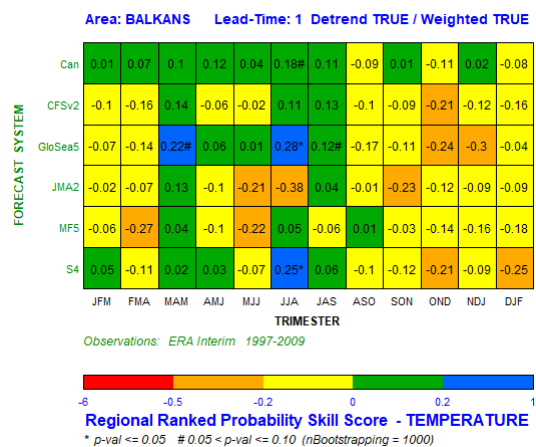
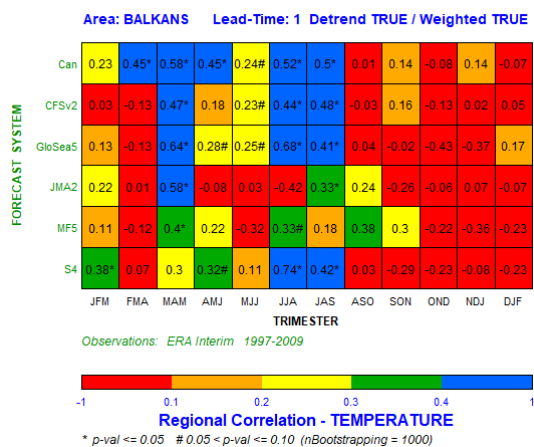


Table 62. The same as Table 60, but for BALKANS domain

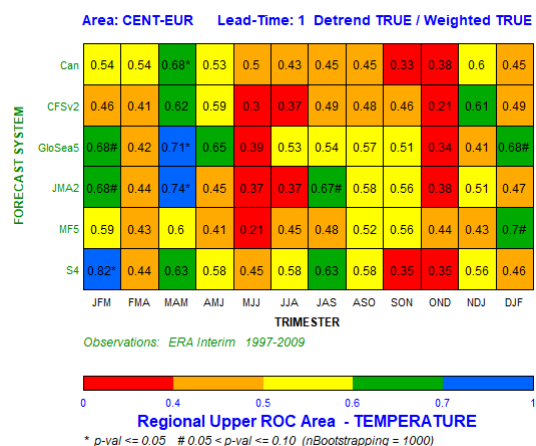
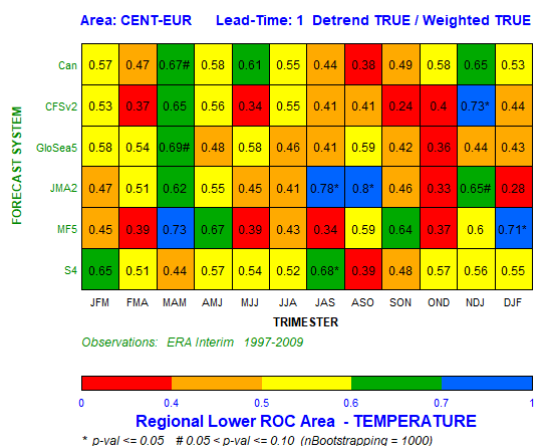
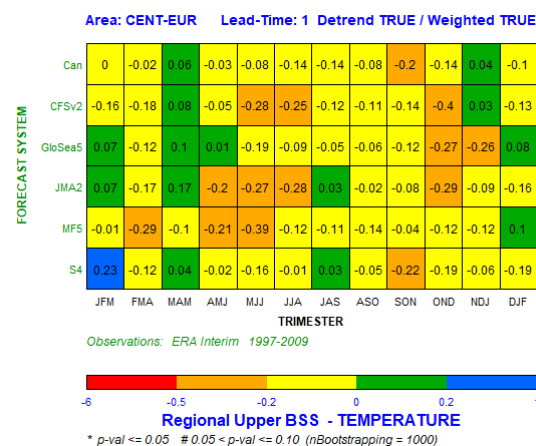
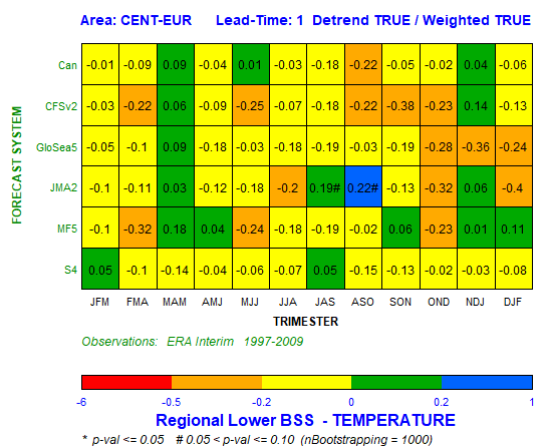
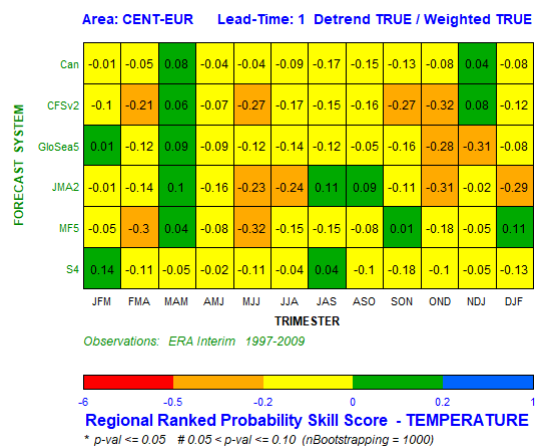
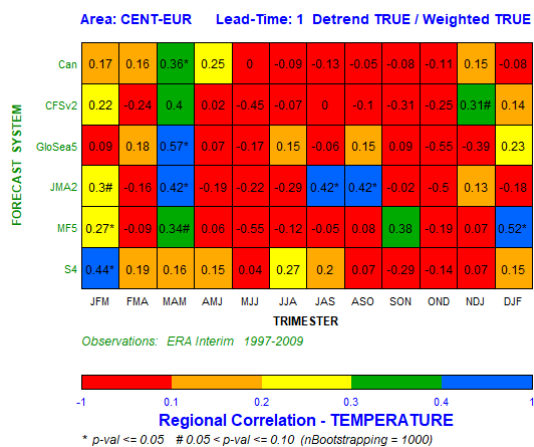


Table 63. The same as Table 60, but for CENTRE EUROPE domain

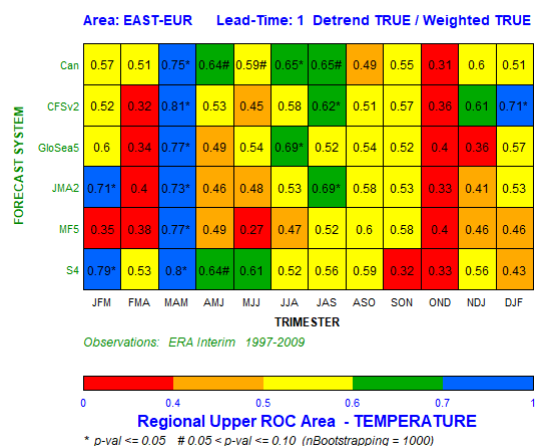
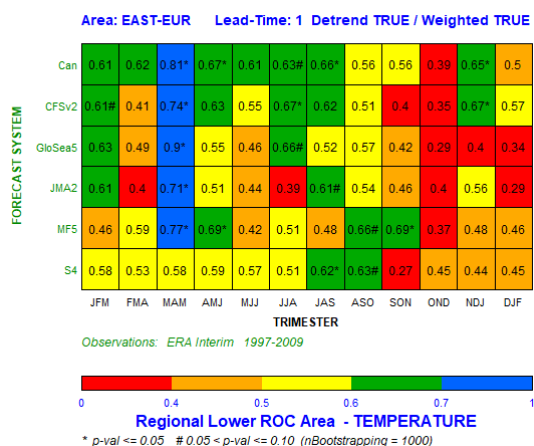
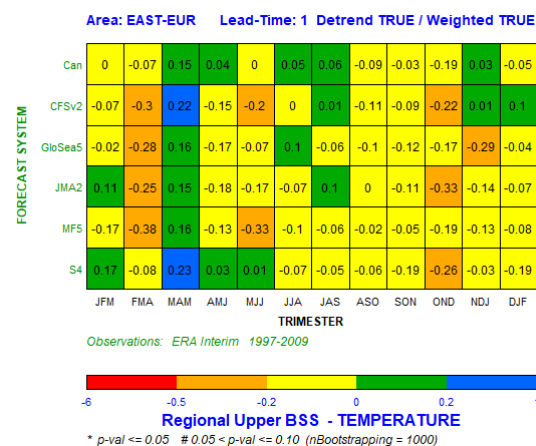
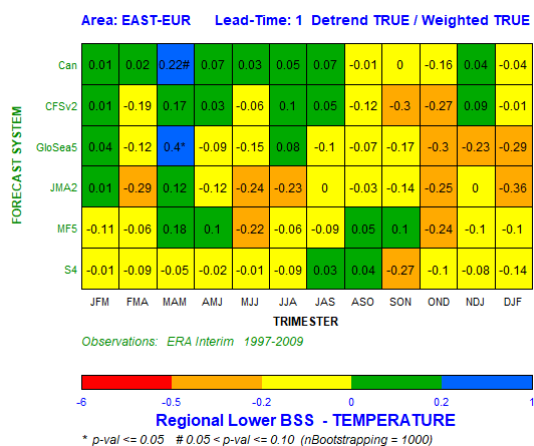
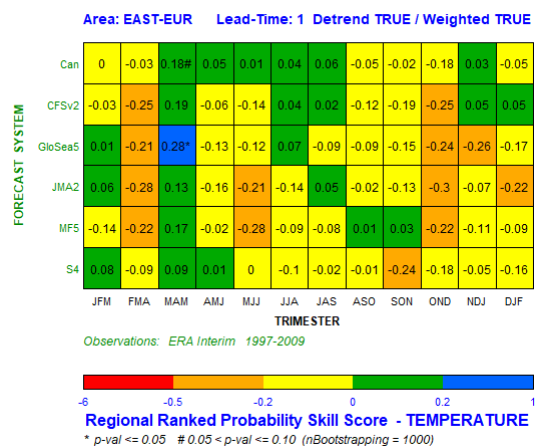
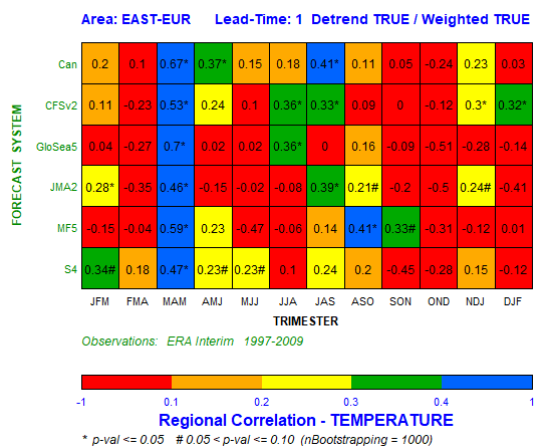


Table 64. The same as Table 60, but for EAST EUROPE domain

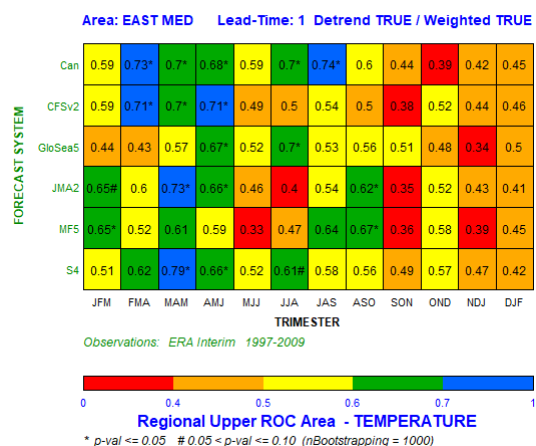
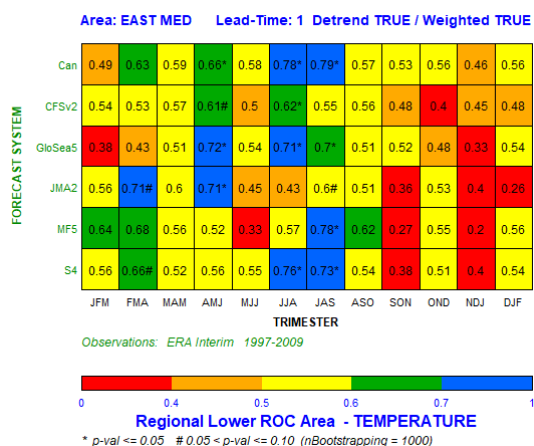
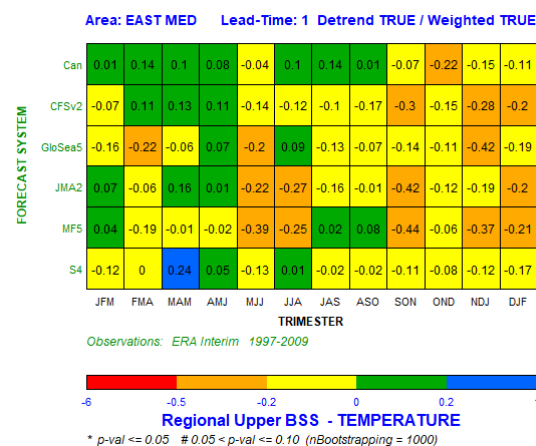
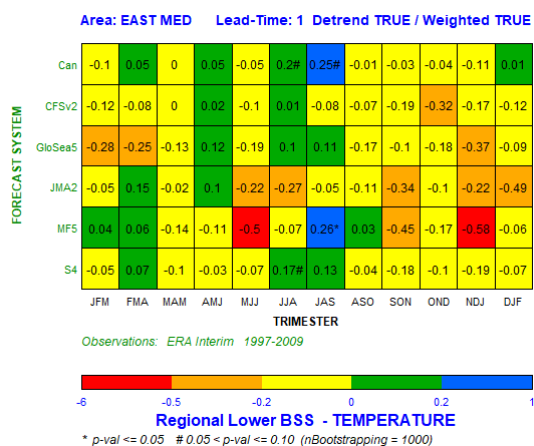
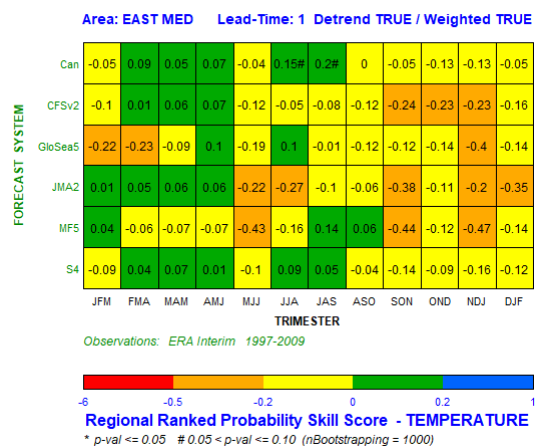
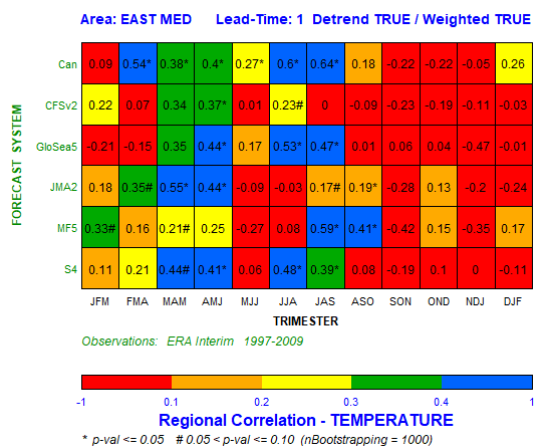


Table 65. The same as Table 60, but for EAST MEDITERRANEAN domain

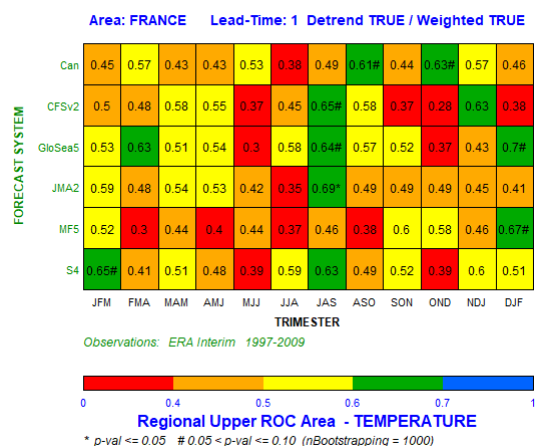
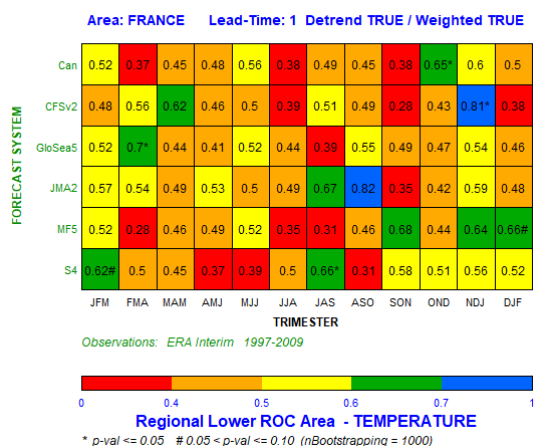
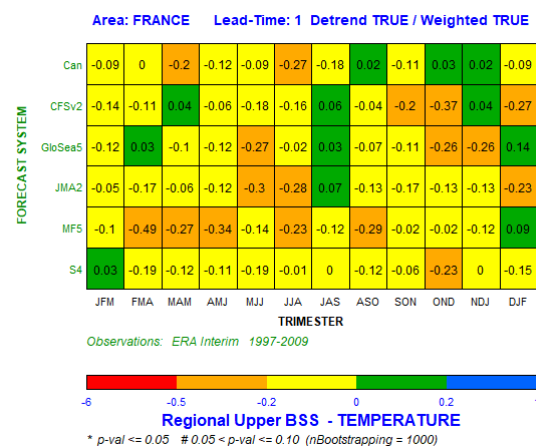
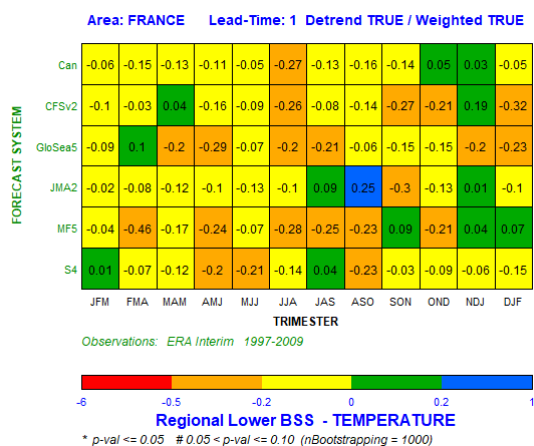
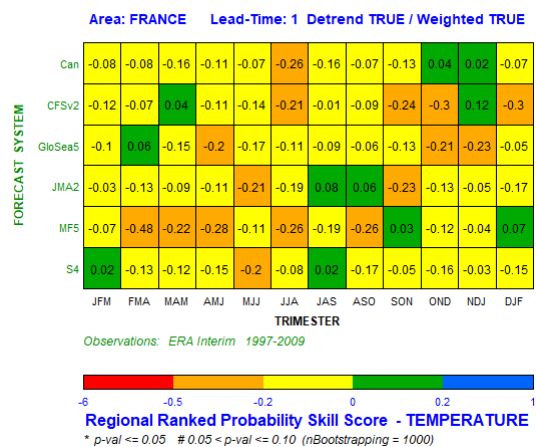
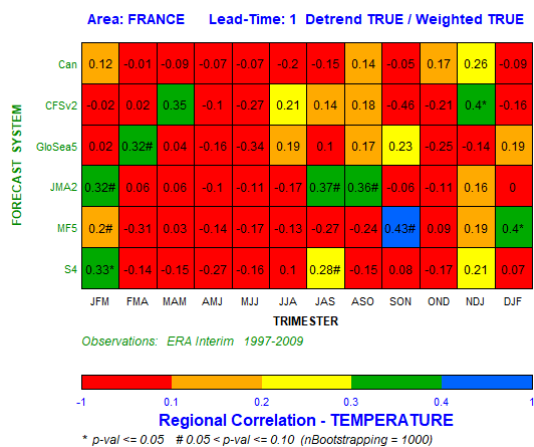


Table 66. The same as Table 60, but for FRANCE domain

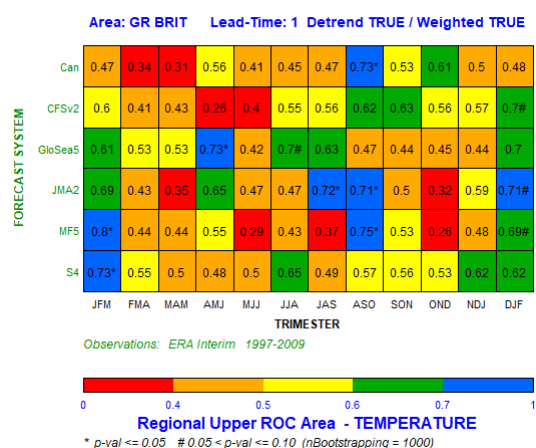
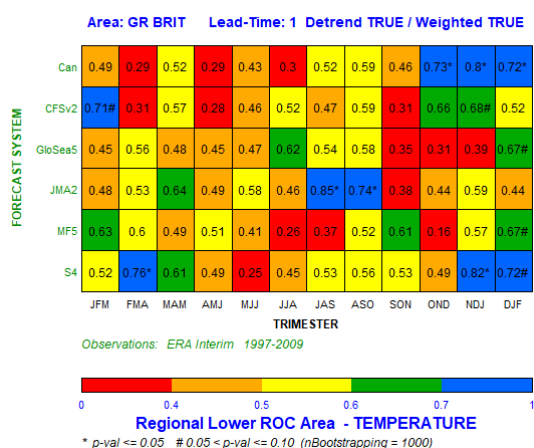
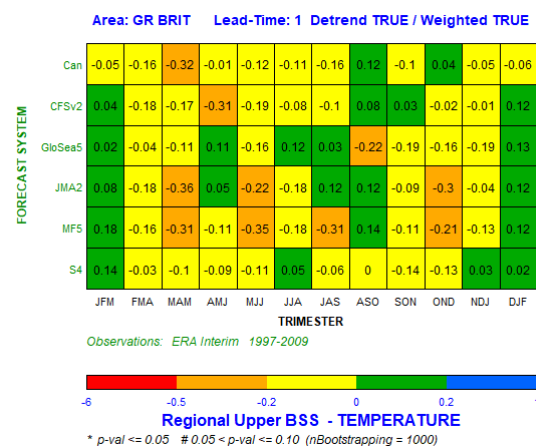
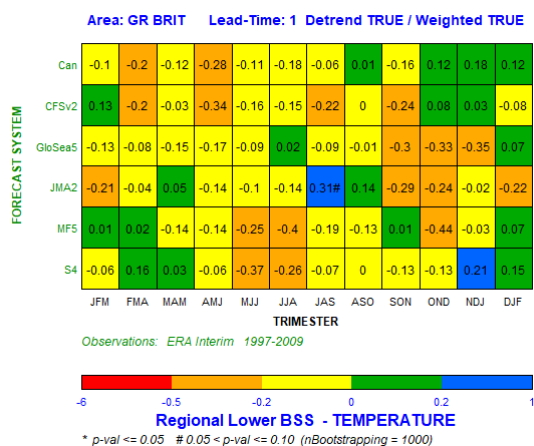
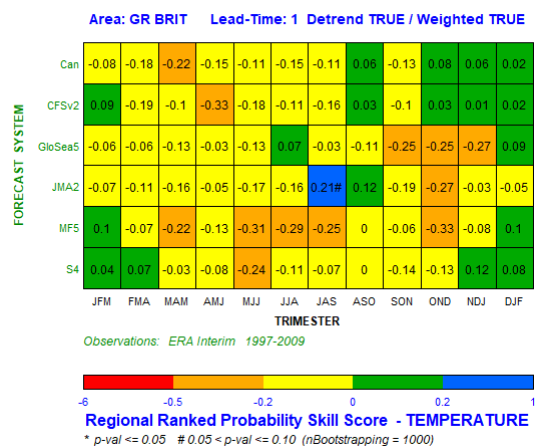
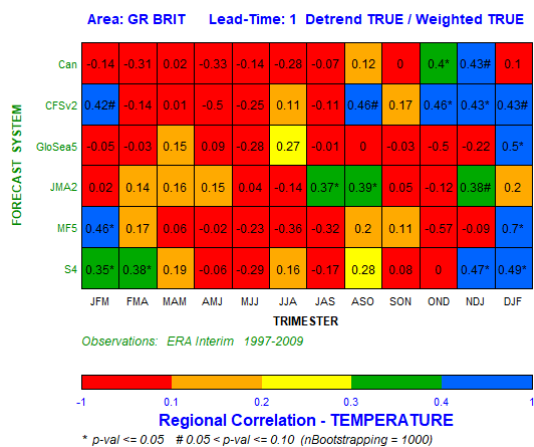


Table 67. The same as Table 60, but for GREAT BRITAIN domain

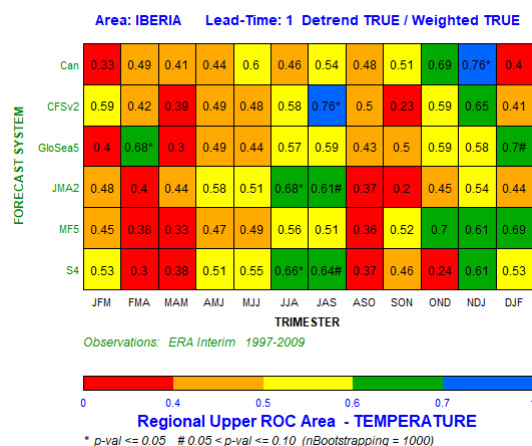
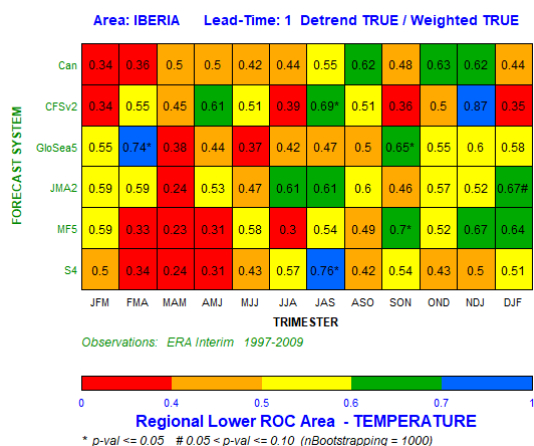
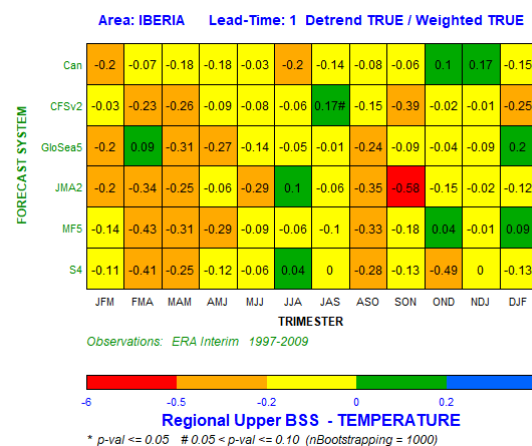
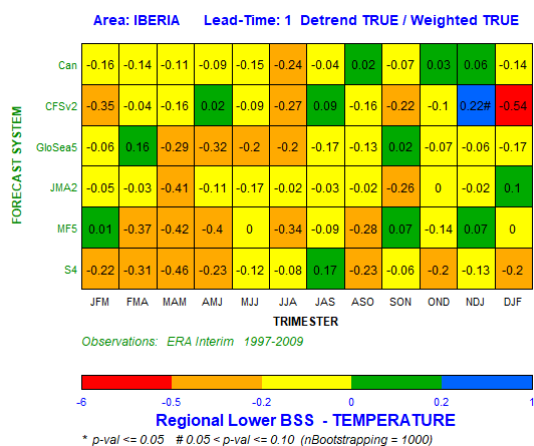
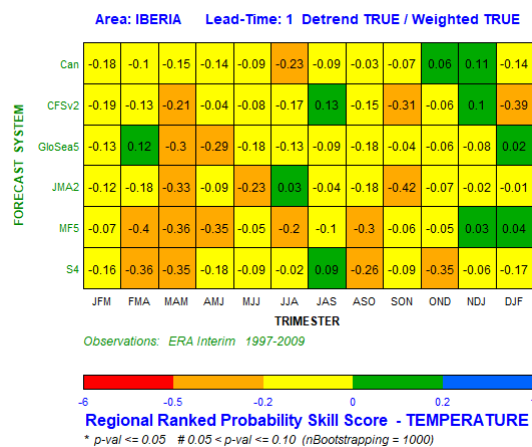
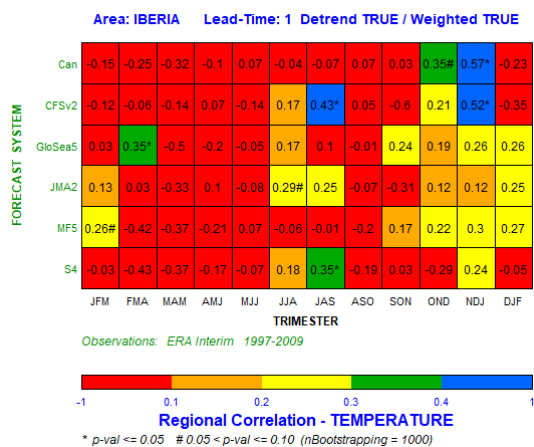


Table 68. The same as Table 60, but for IBERIA domain

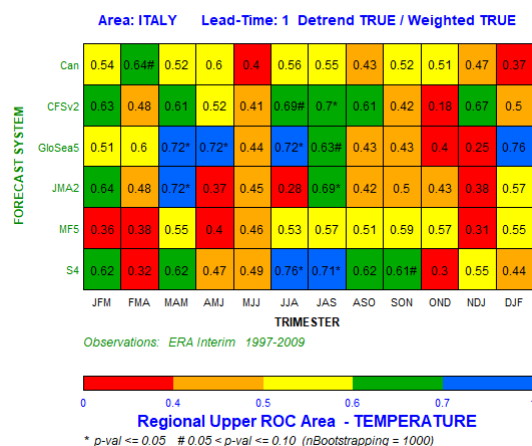
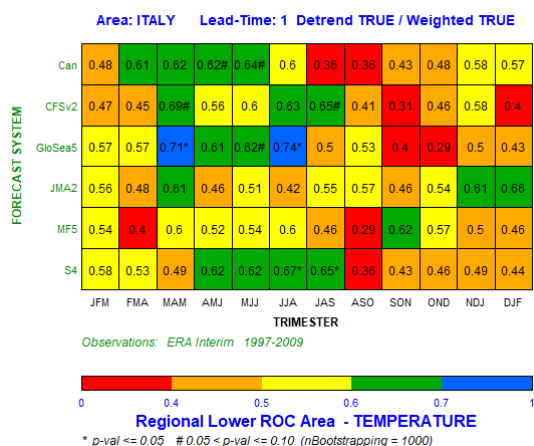
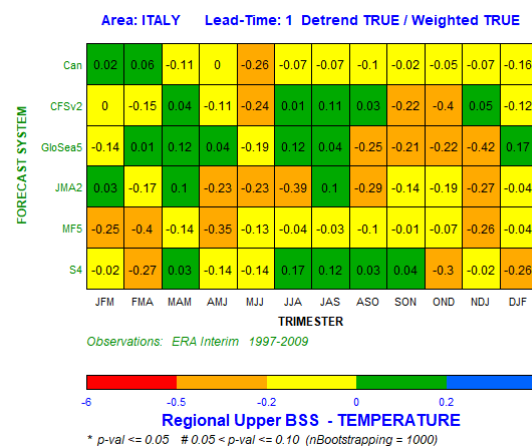
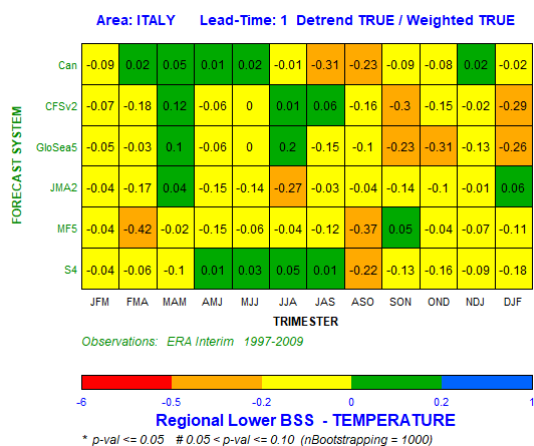
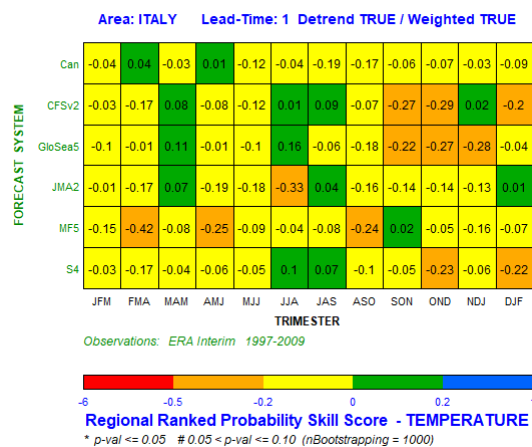
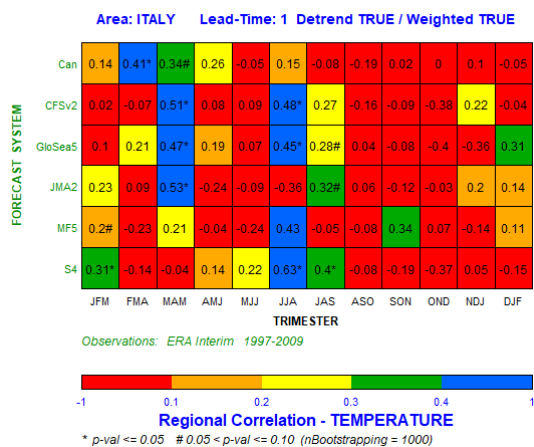


Table 69. The same as Table 60, but for ITALY domain

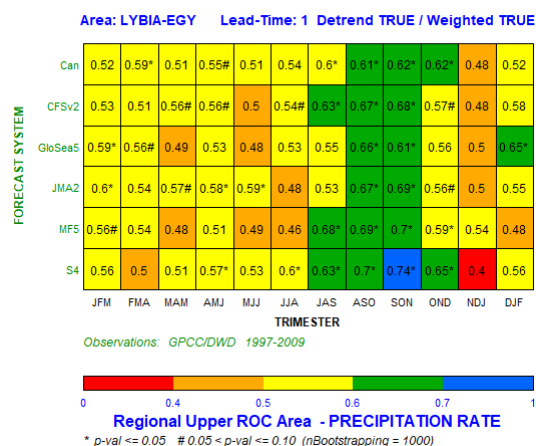
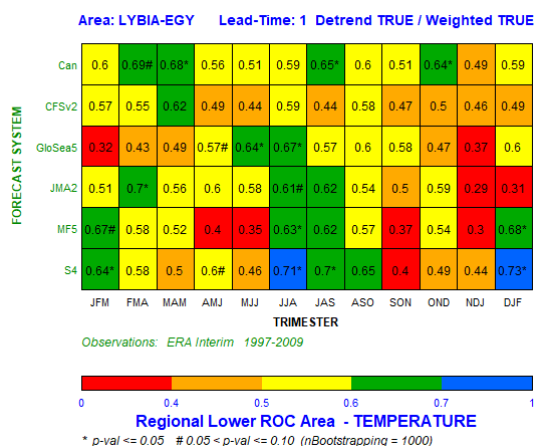
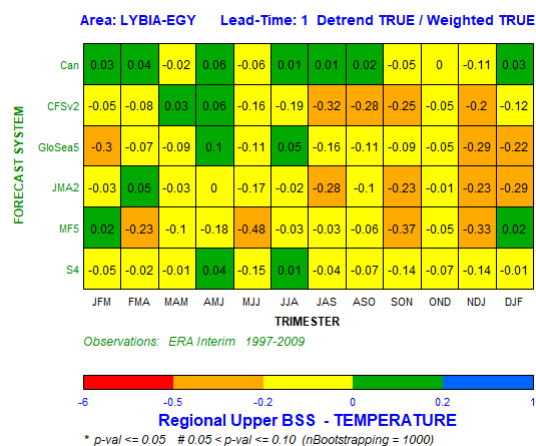
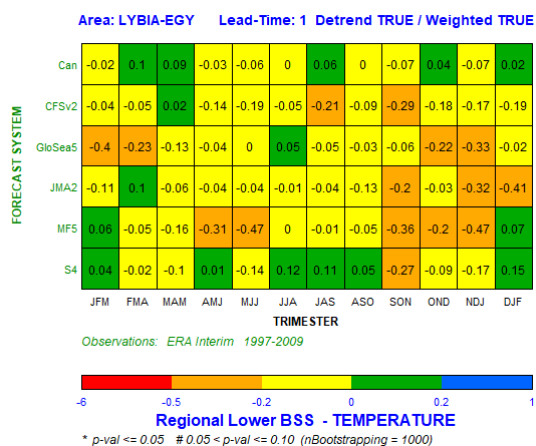
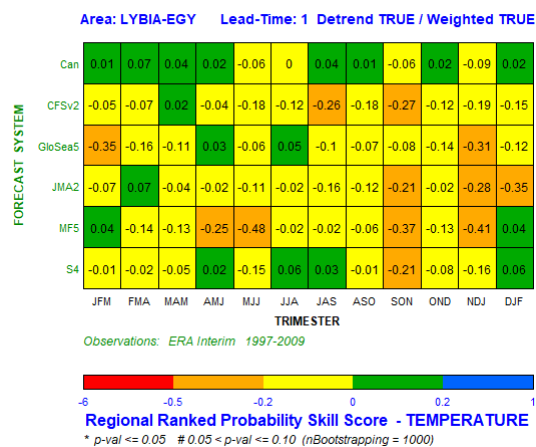
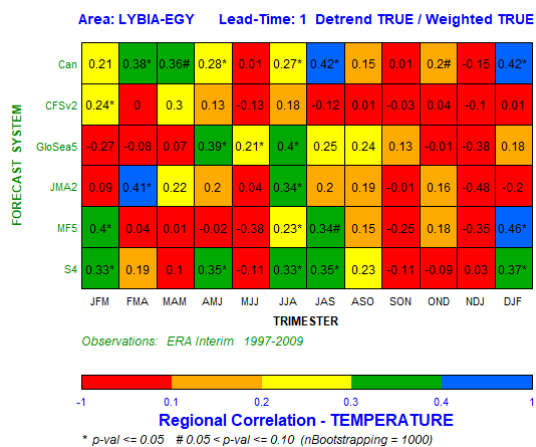


Table 70. The same as Table 60, but for LIBYA-EGYPT domain

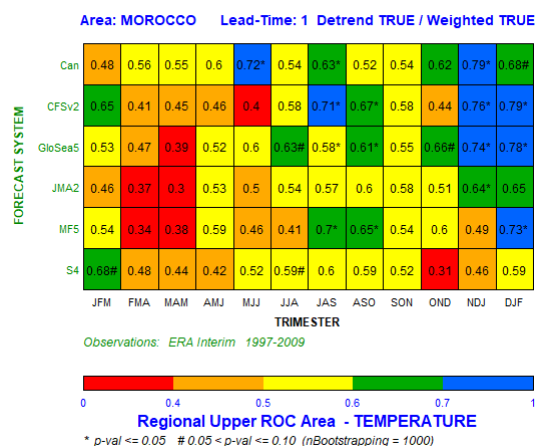
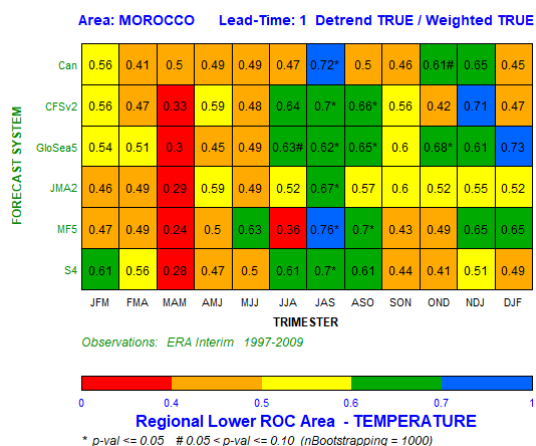
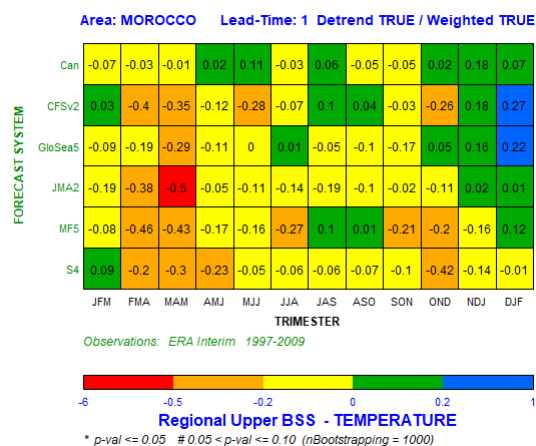
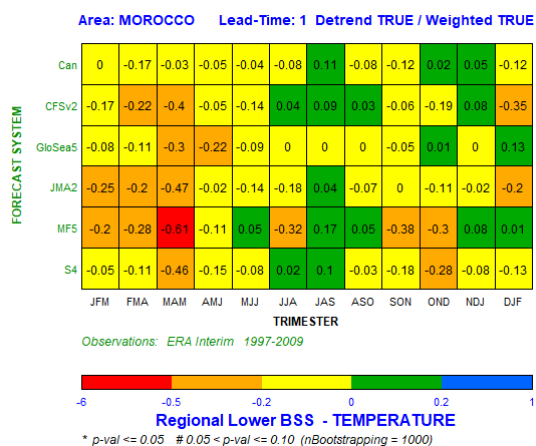
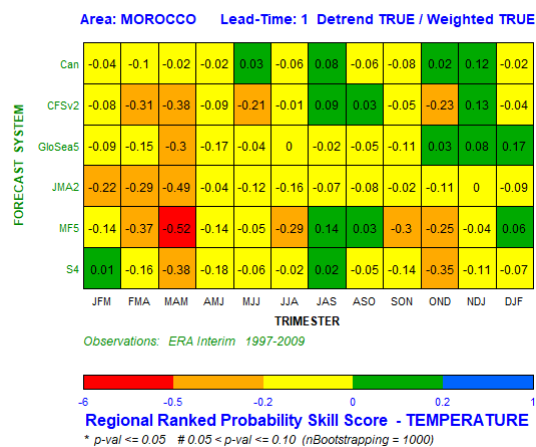
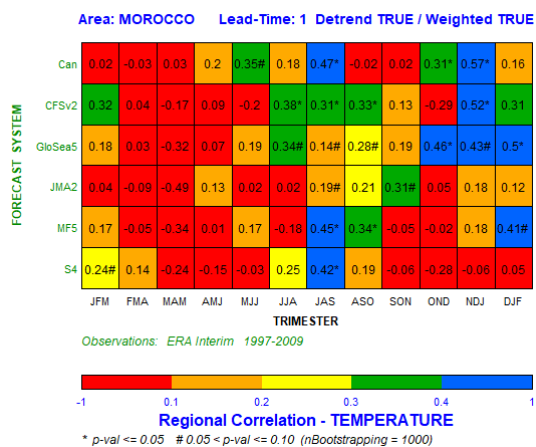


Table 71. The same as Table 60, but for MOROCCO domain

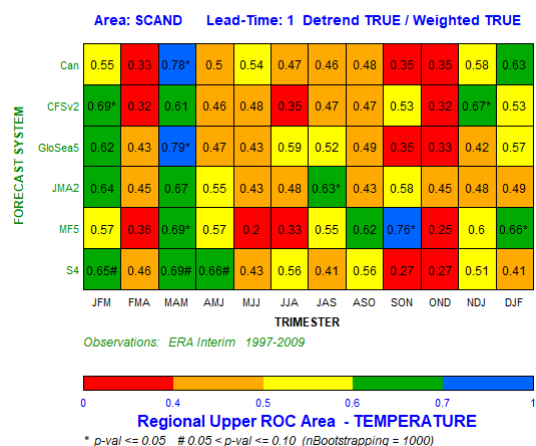
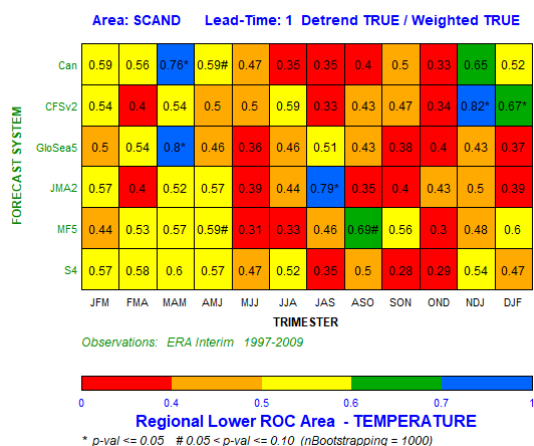
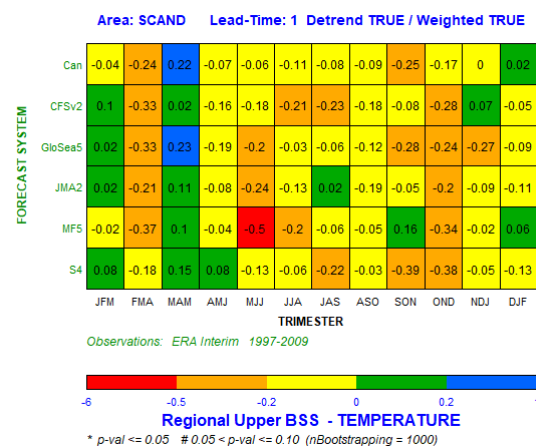
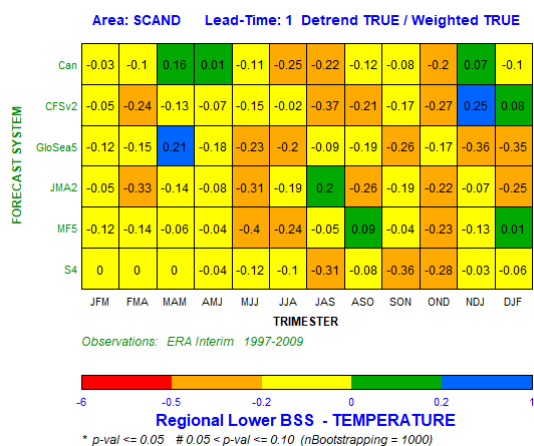
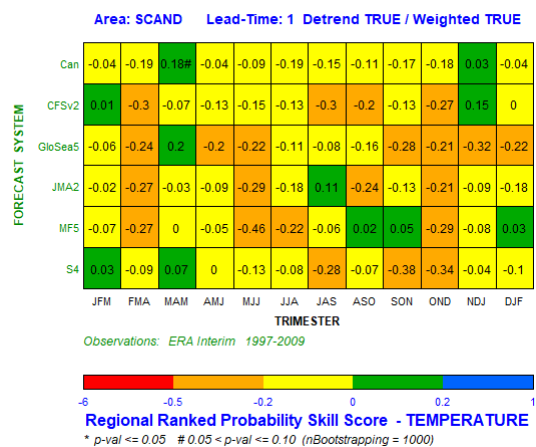
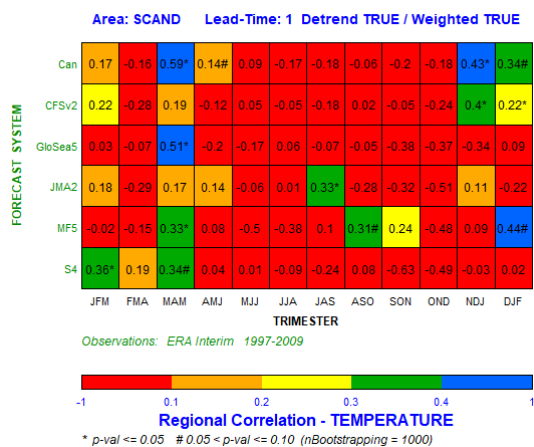


Table 72. The same as Table 60, but for SCANDINAVIA domain

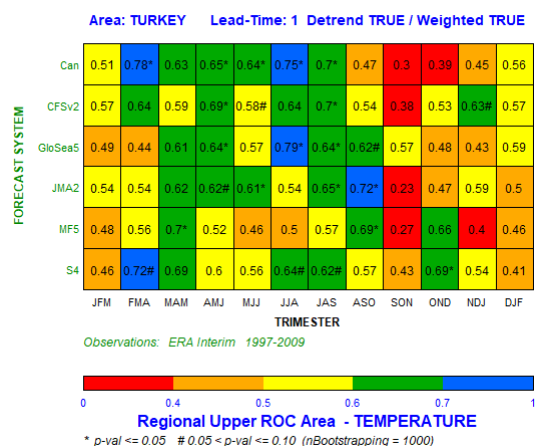
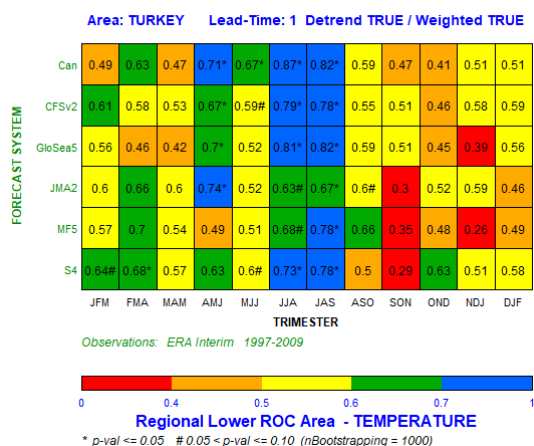
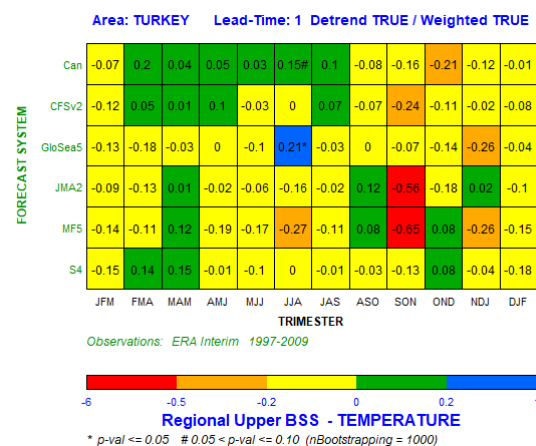
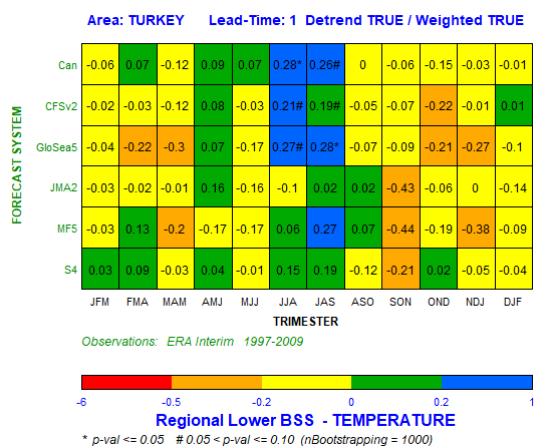
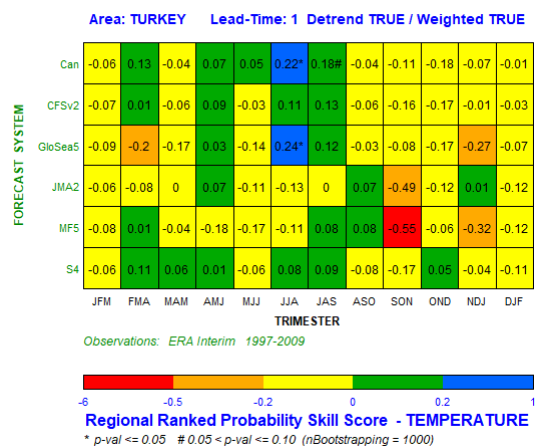
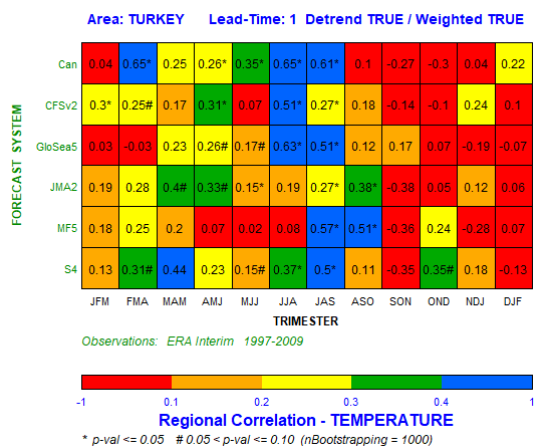


Table 73. The same as Table 60, but for TURKEY domain

2. VERIFICATION PERIOD: Maximum available period for each model

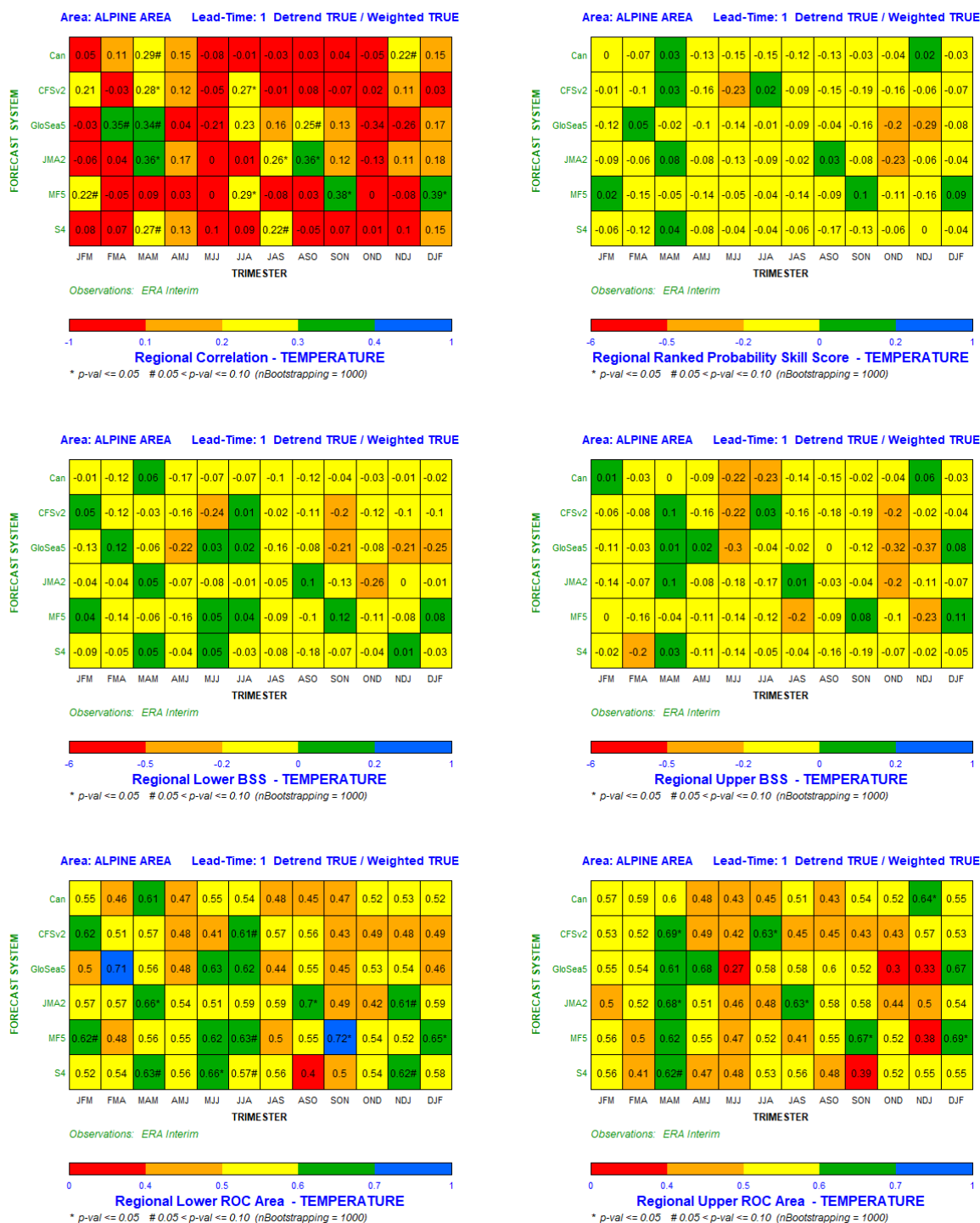


Table 74. Regional correlation coefficient, Ranked Probability Skill Score, Lower and Upper ROC Area and BSS computed for the anomaly values of temperature removing trend, for the 12 different three-month periods and for lead-time 1 over ALPINE AREA domain for the maximum available period for each model. The three-month periods for the seasonal forecasts are done, moved one by one for each column in the table, are shown in the X-axis. The direct outputs of the different models (S4, MF5, JMA2, GLOSEA5, CFSv2 and Can) are represented in the Y-axis (see text for their description).

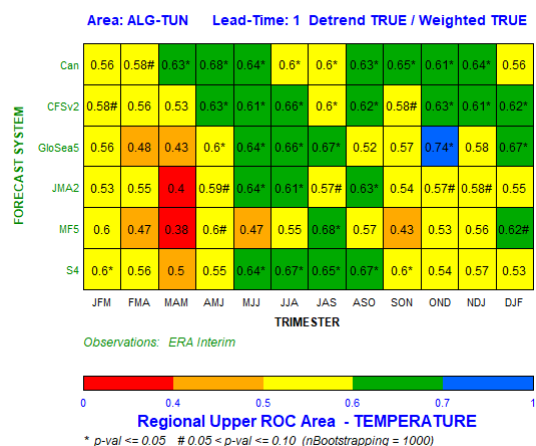
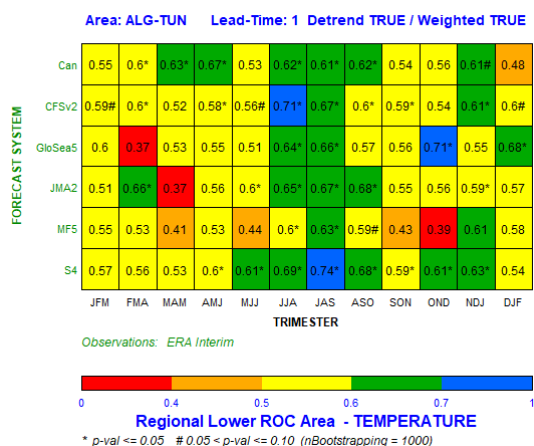
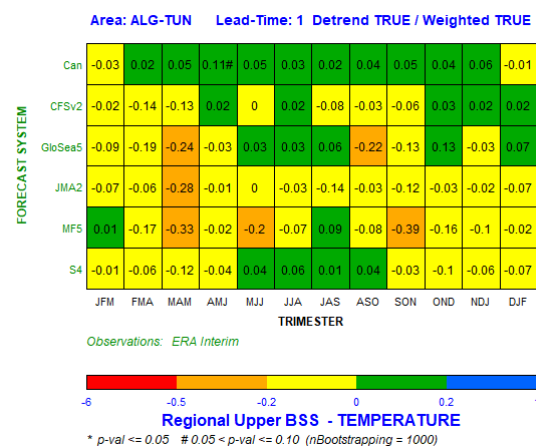
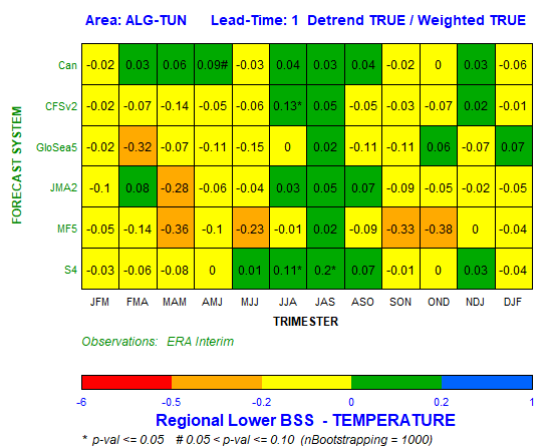
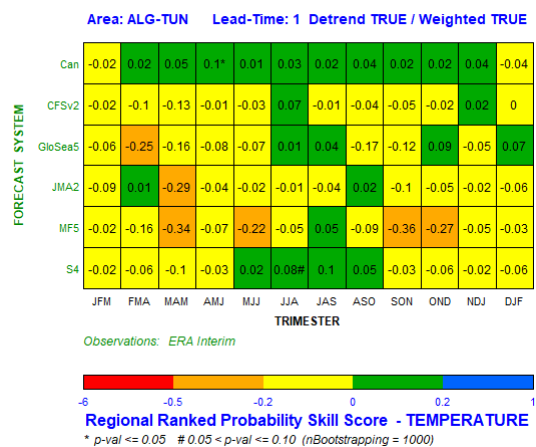
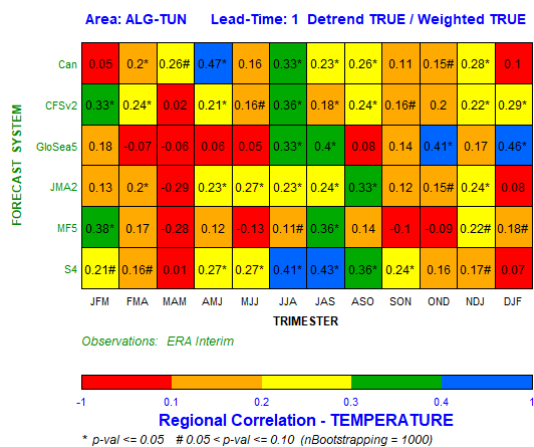


Table 75. The same as Table 74, but for ALGERIA/TUNISIA domain

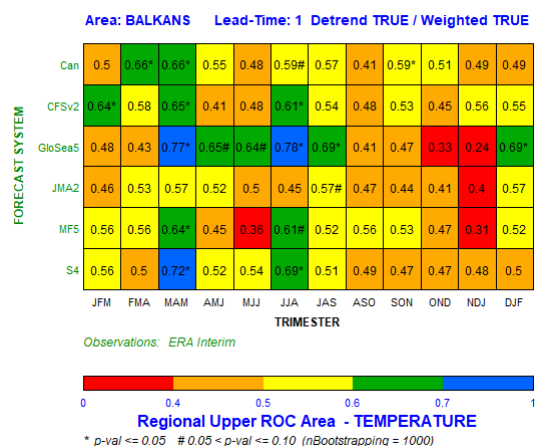
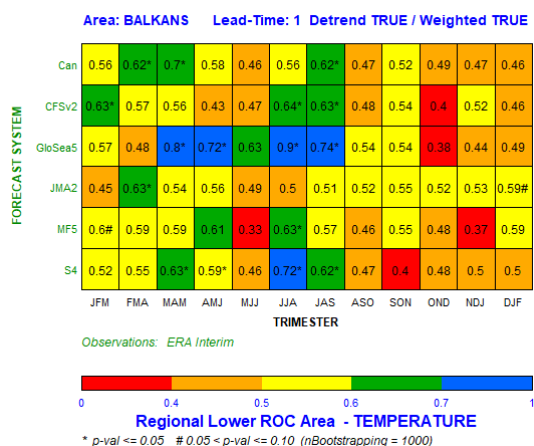
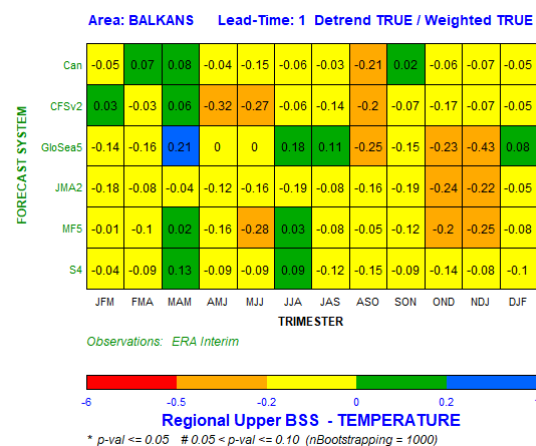
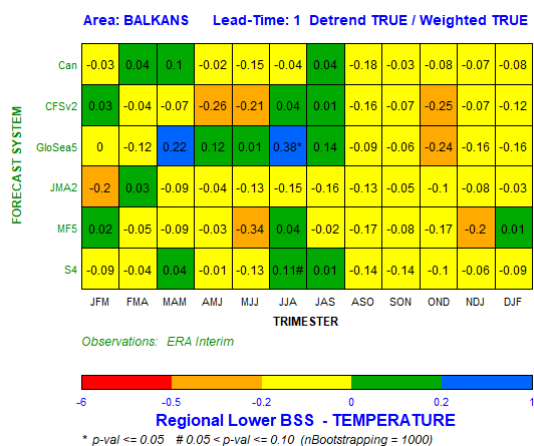
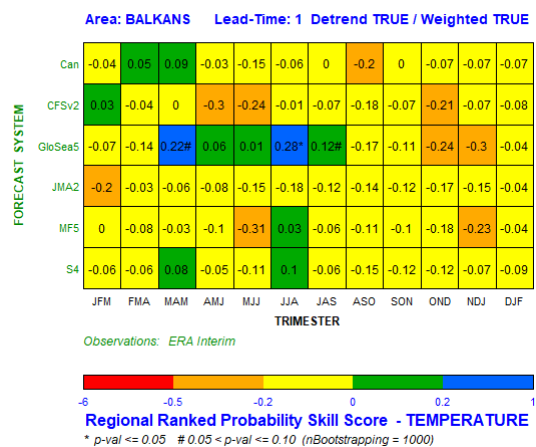
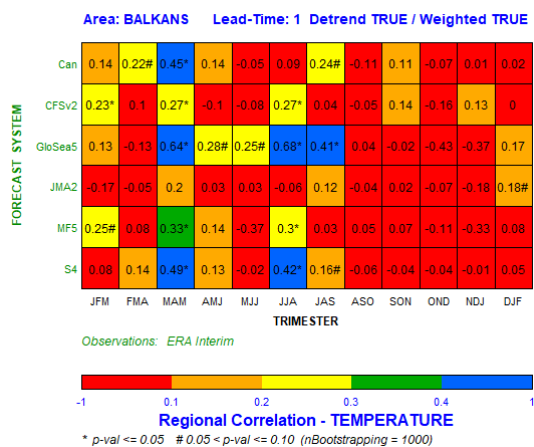


Table 76. The same as Table 74, but for BALKANS domain

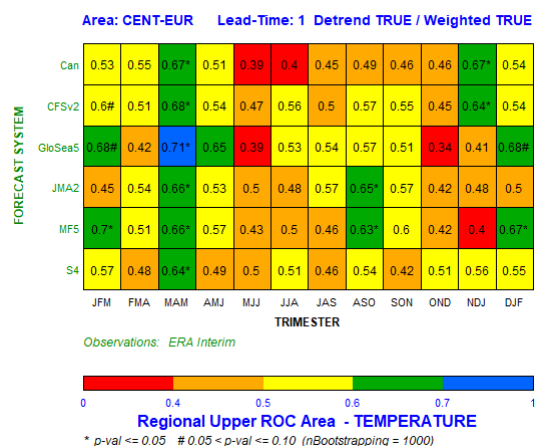
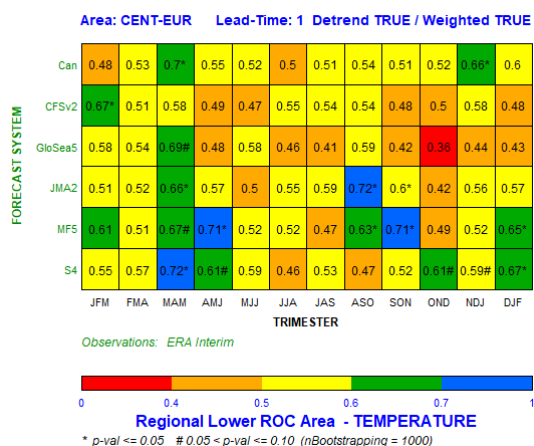
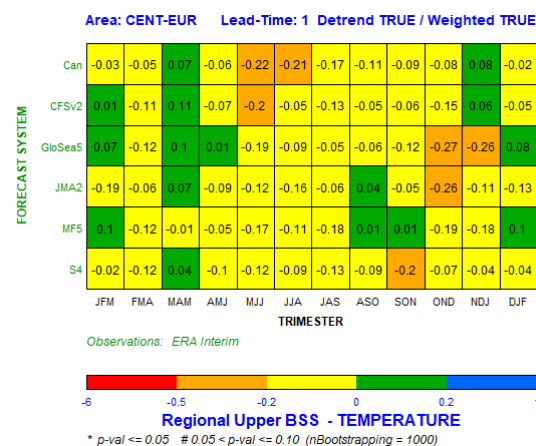
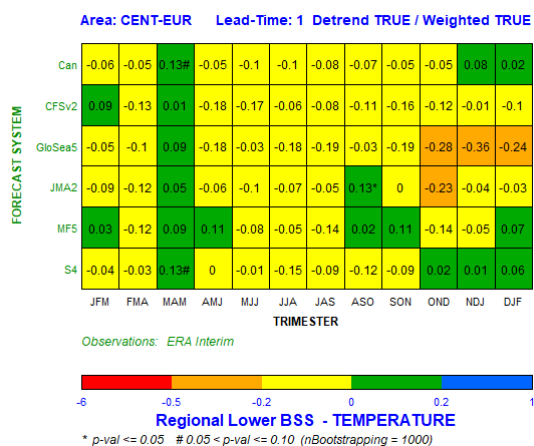
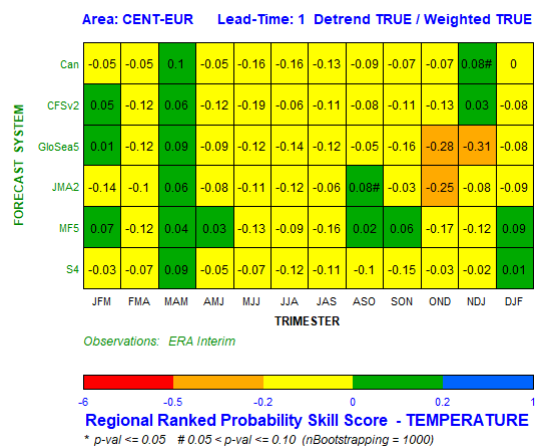
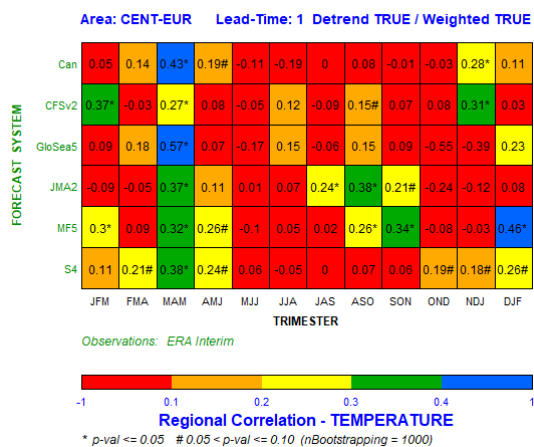


Table 77. The same as Table 74, but for CENTRE EUROPE domain

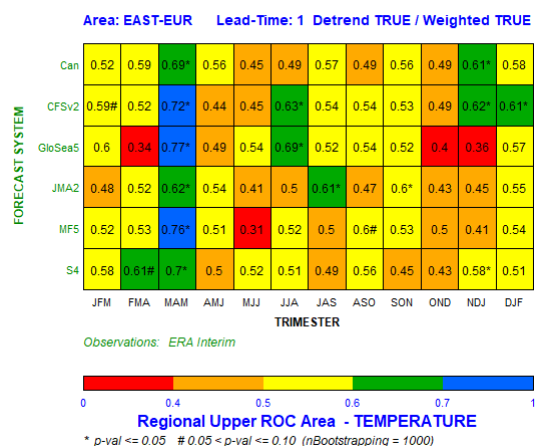
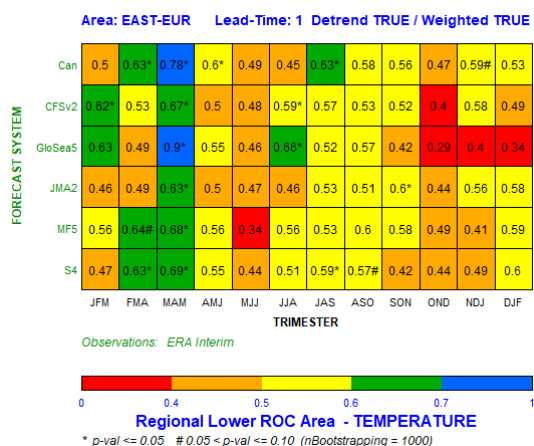
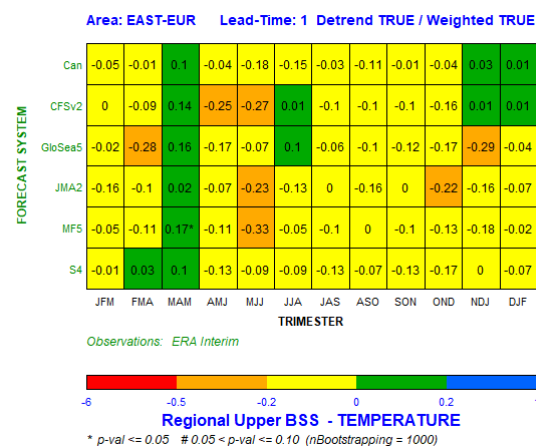
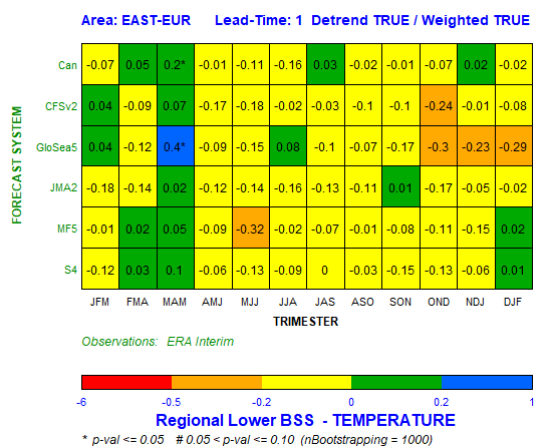
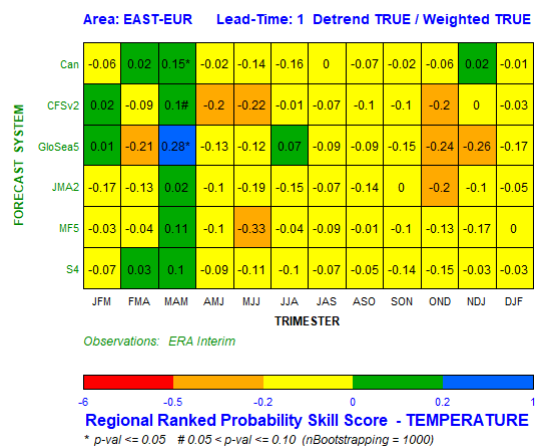
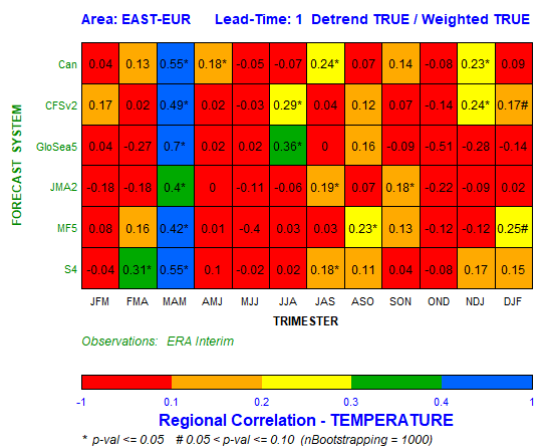


Table 78. The same as Table 74, but for EAST EUROPE domain

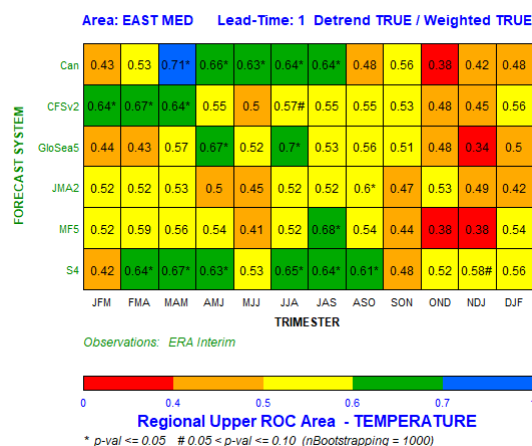
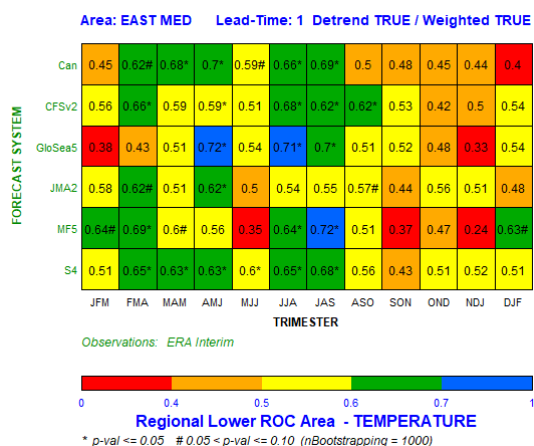
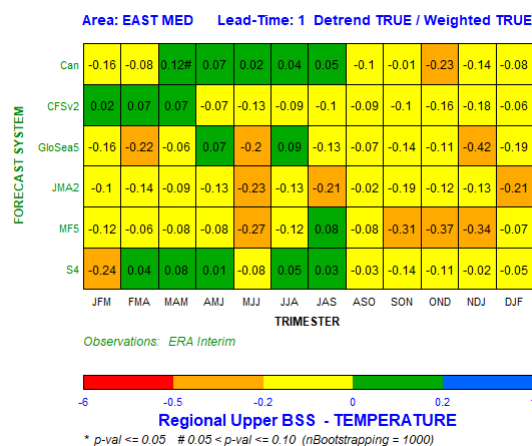
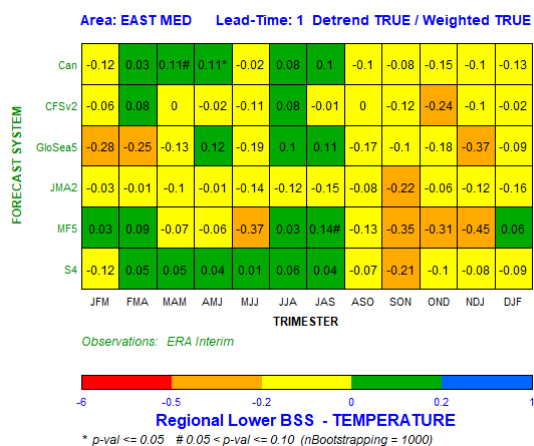
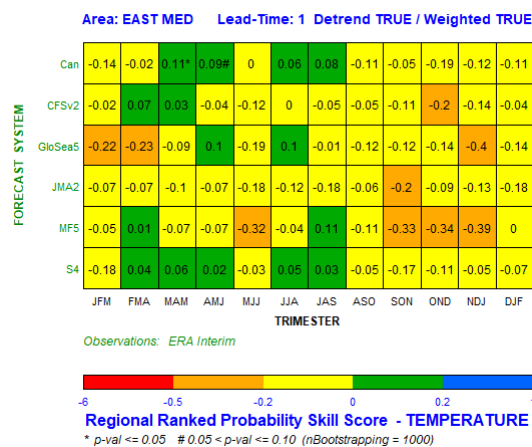
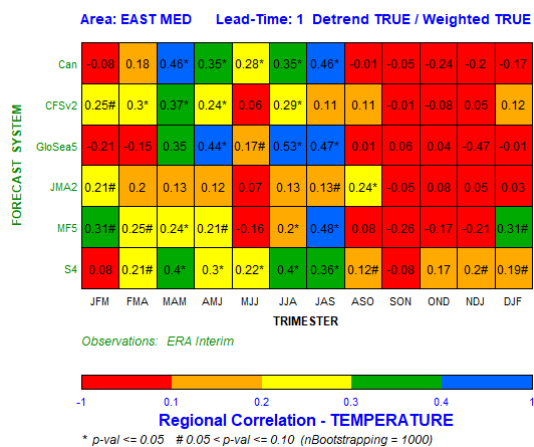


Table 79. The same as Table 74, but for EAST MEDITERRANEAN domain

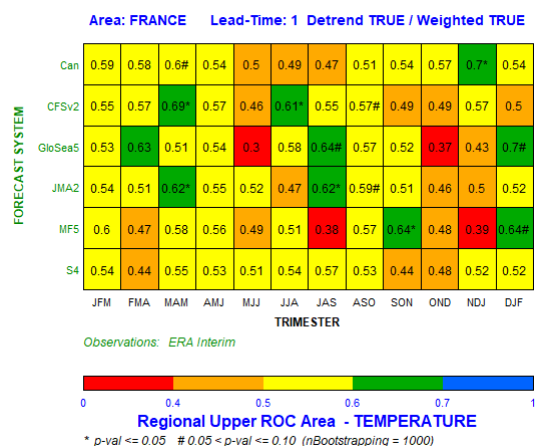
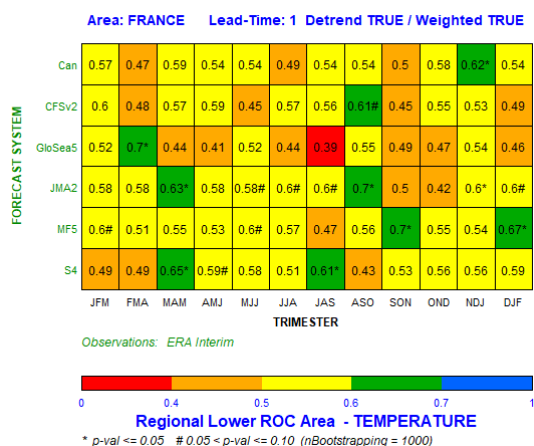
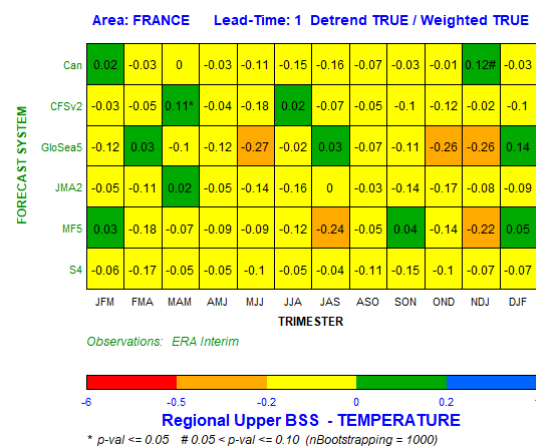
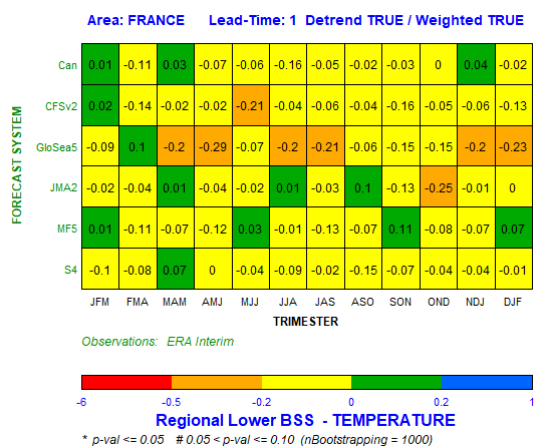
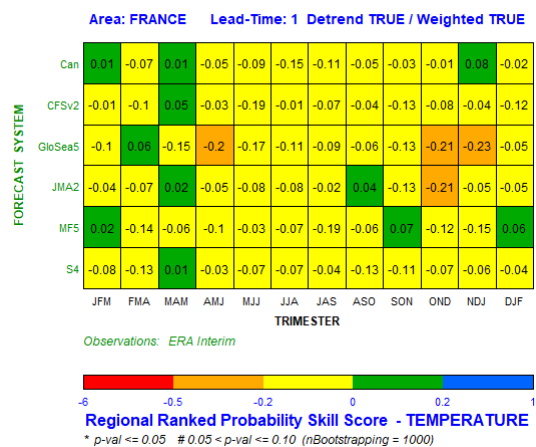
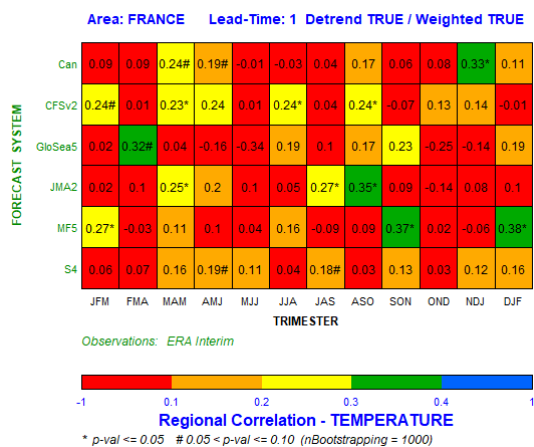


Table 80. The same as Table 74, but for FRANCE domain

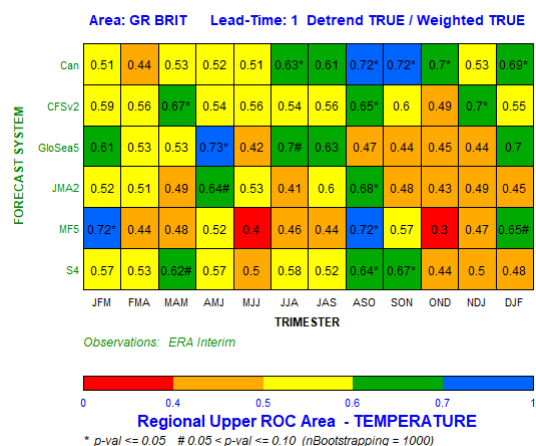
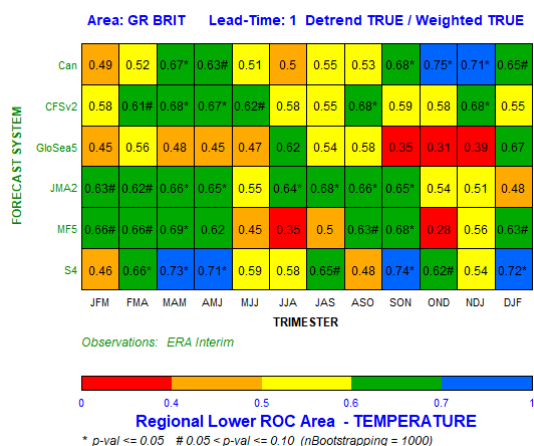
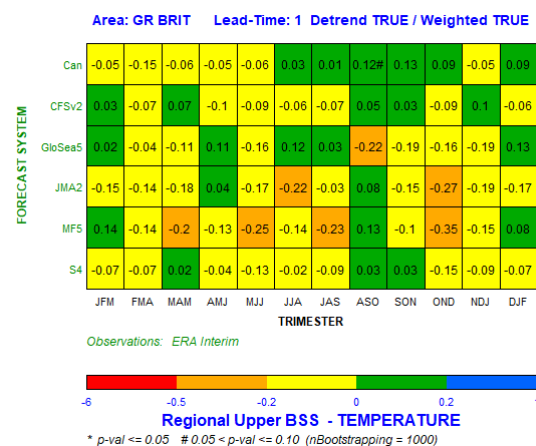
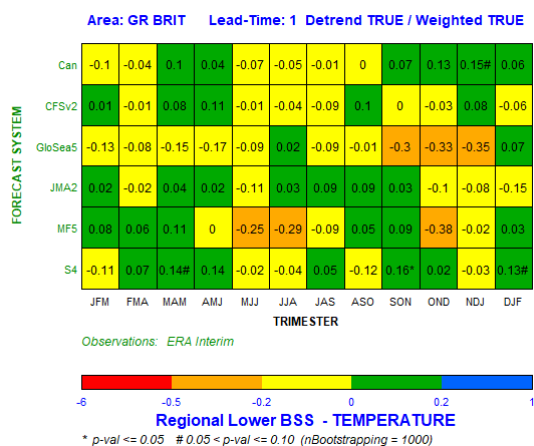
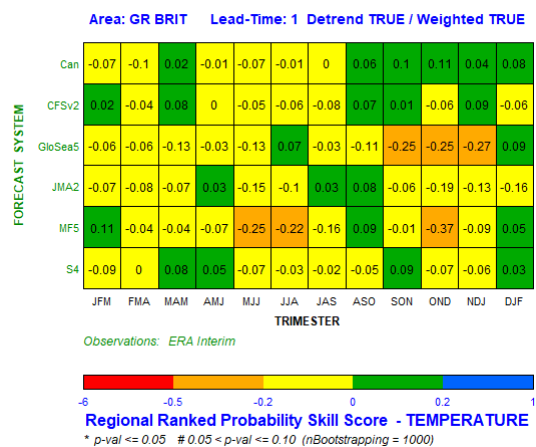
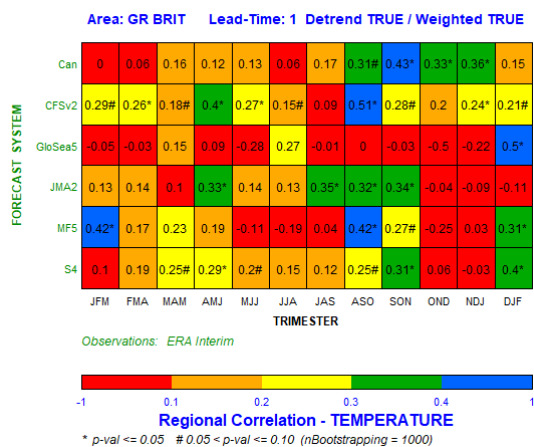


Table 81. The same as Table 74, but for GREAT BRITAIN domain

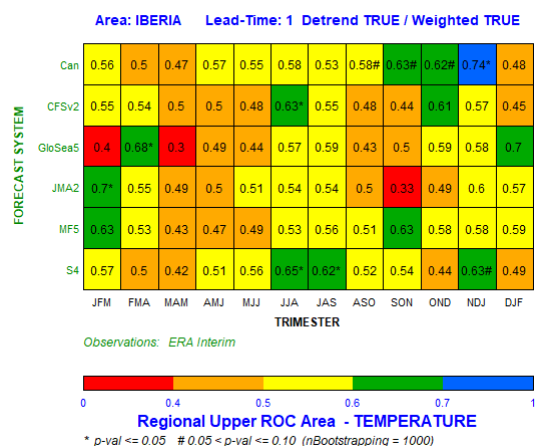
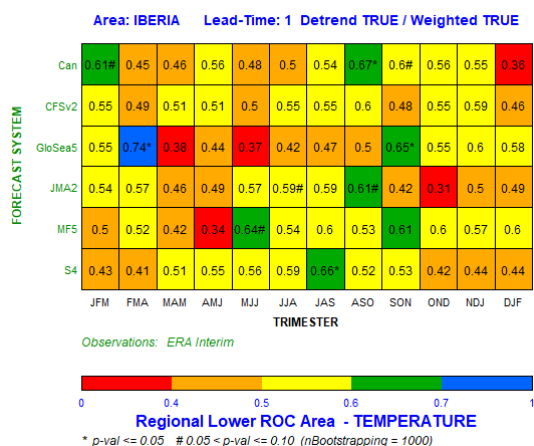
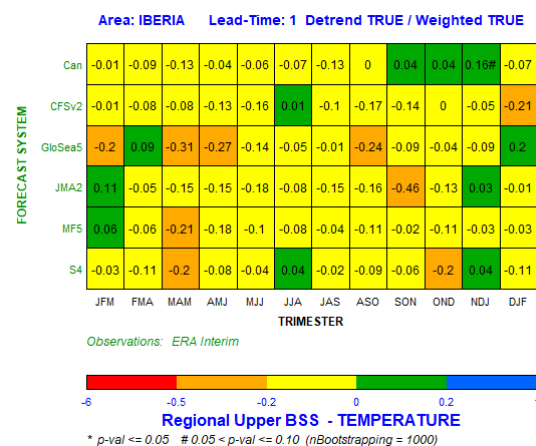
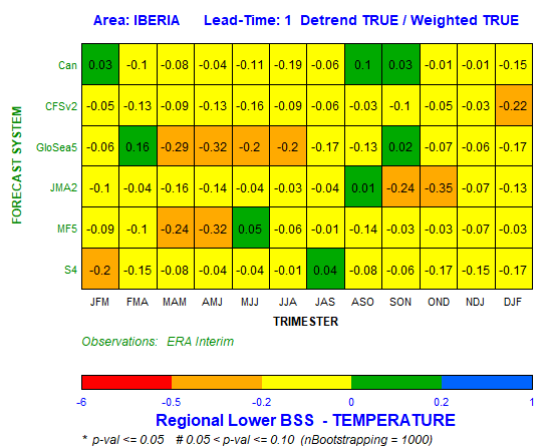
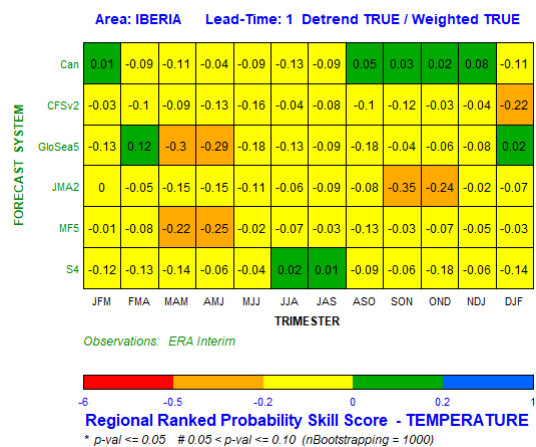
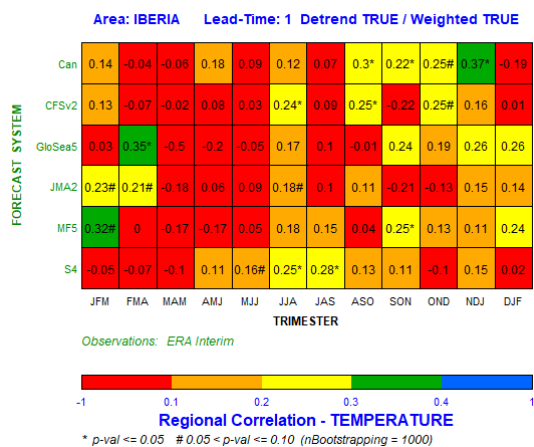


Table 82. The same as Table 74, but for IBERIA domain

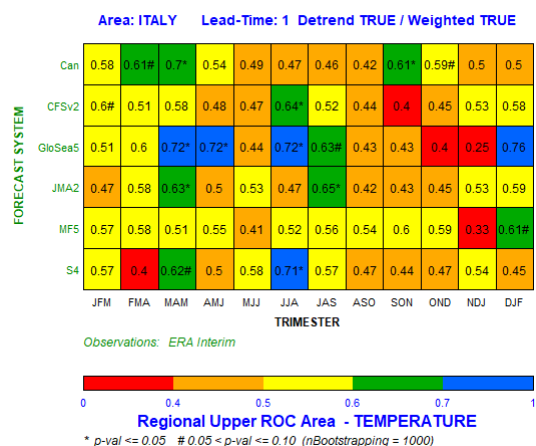
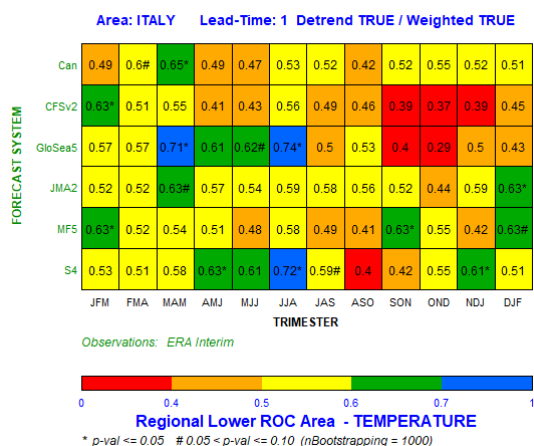
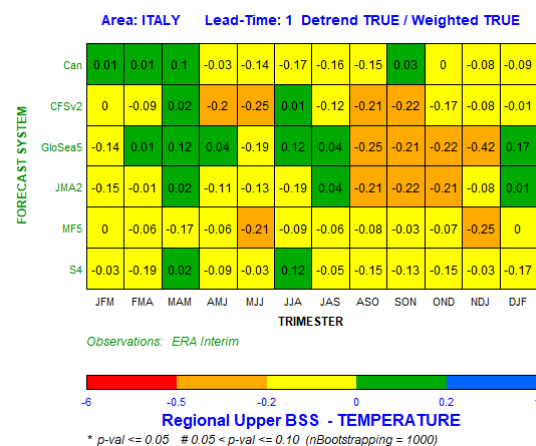
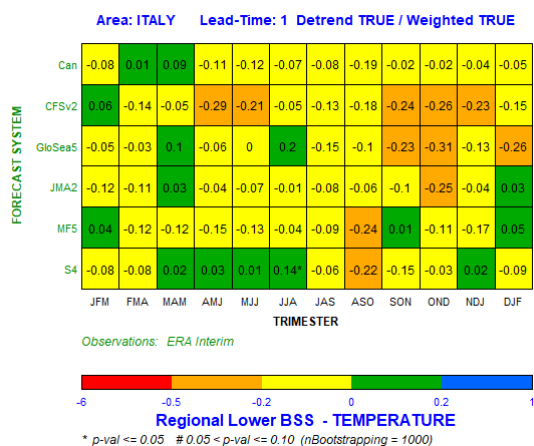
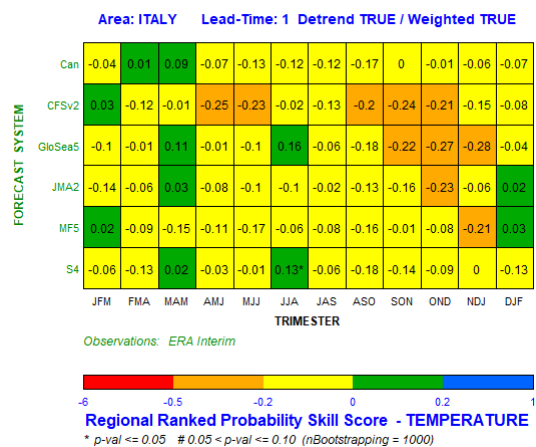
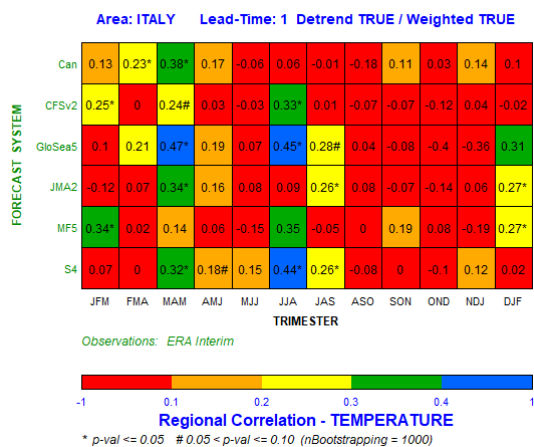


Table 83. The same as Table 74, but for ITALY domain

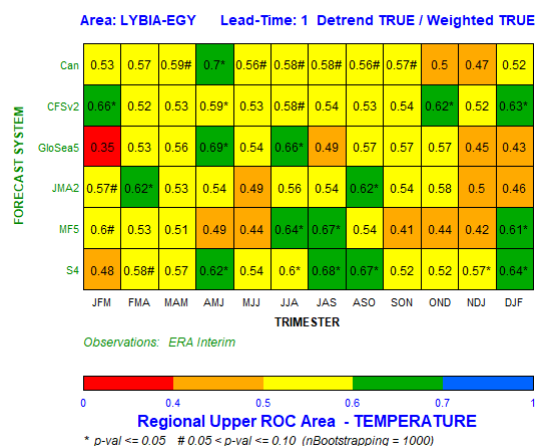
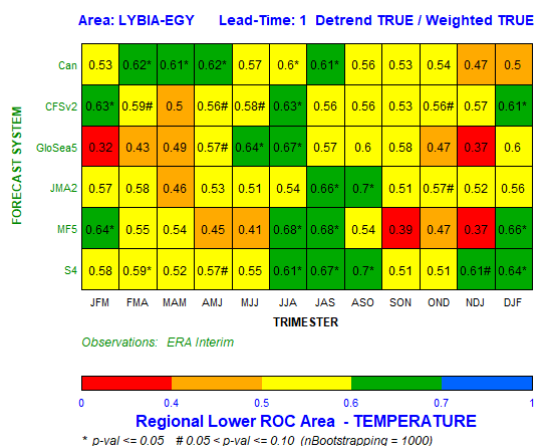
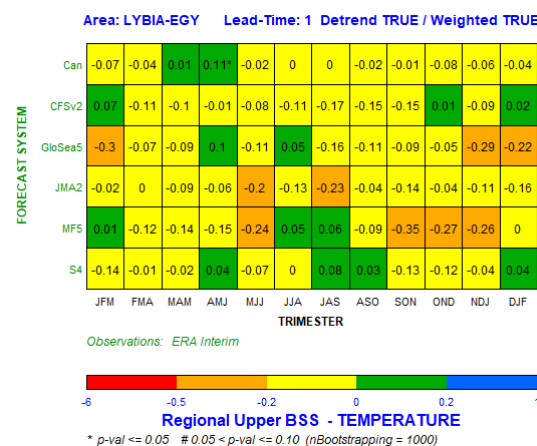
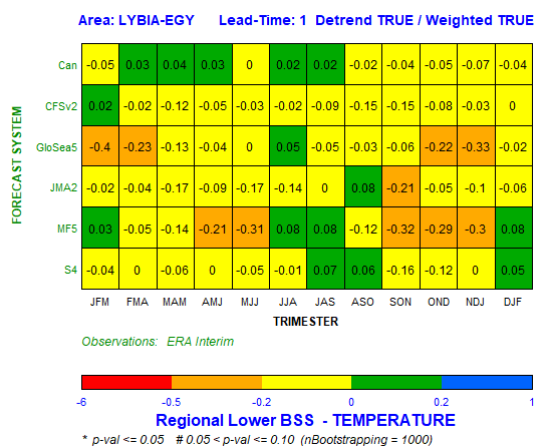
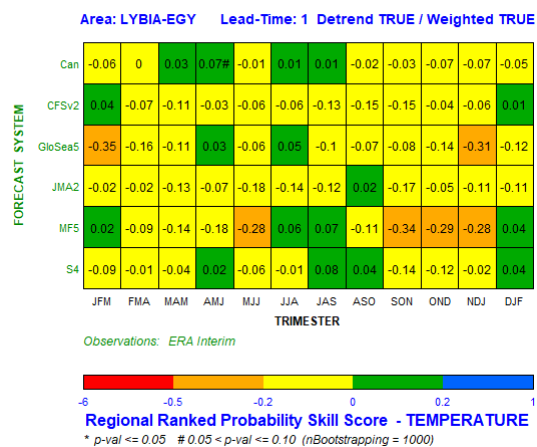
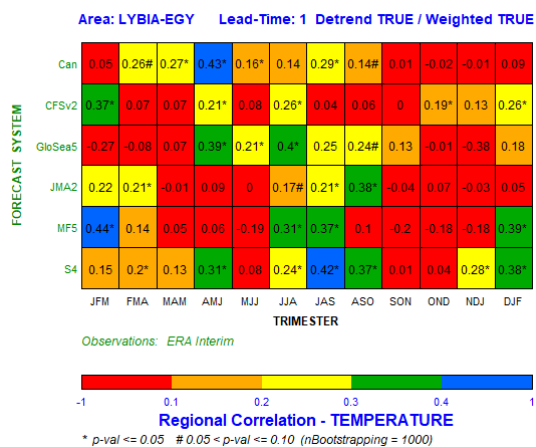


Table 84. The same as Table 74, but for LIBYA-EGYPT domain

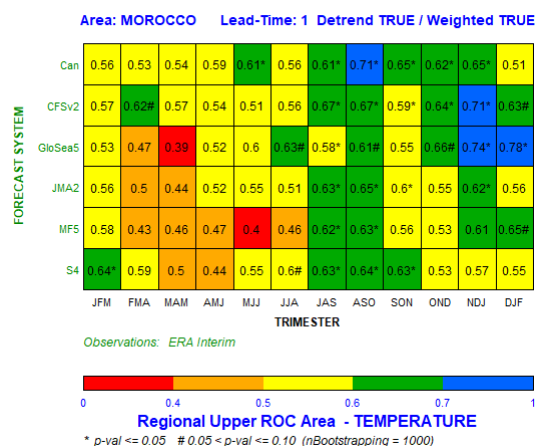
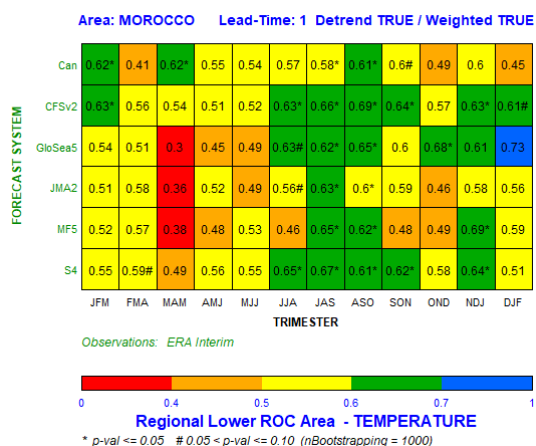
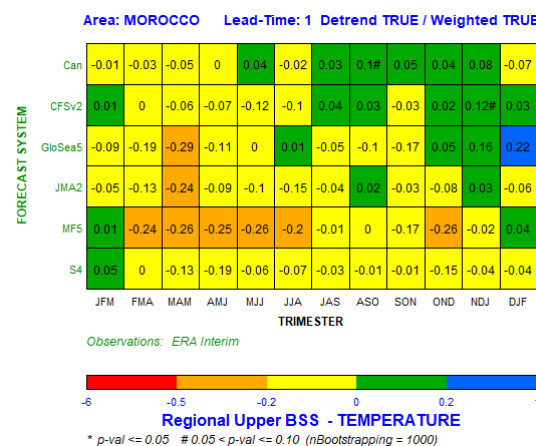
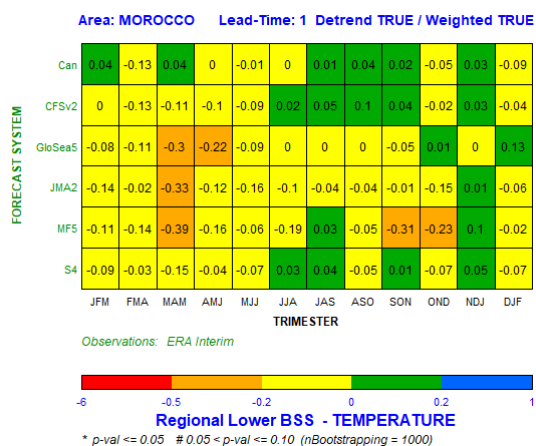
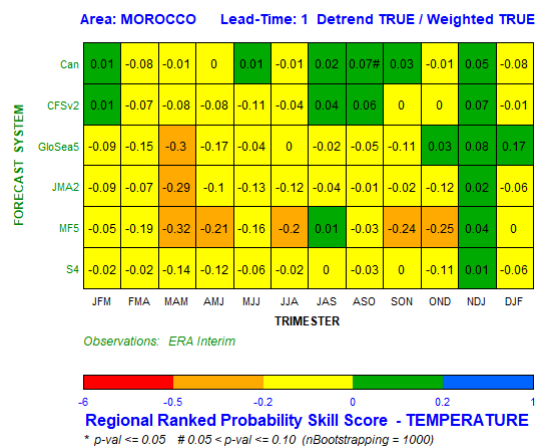
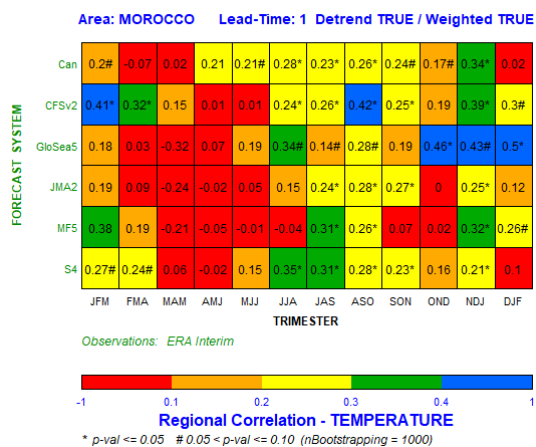


Table 85. The same as Table 74, but for MOROCCO domain

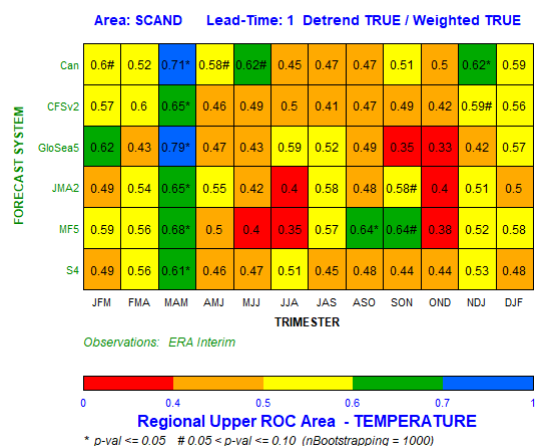
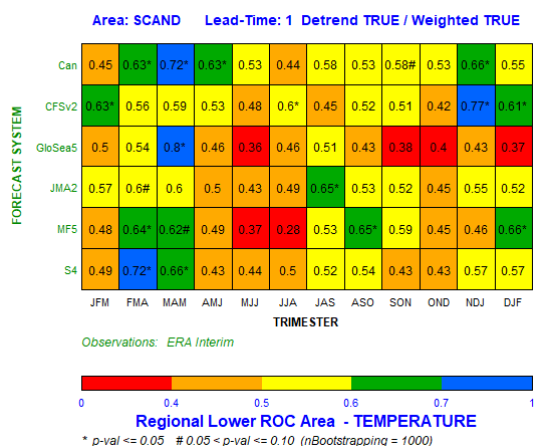
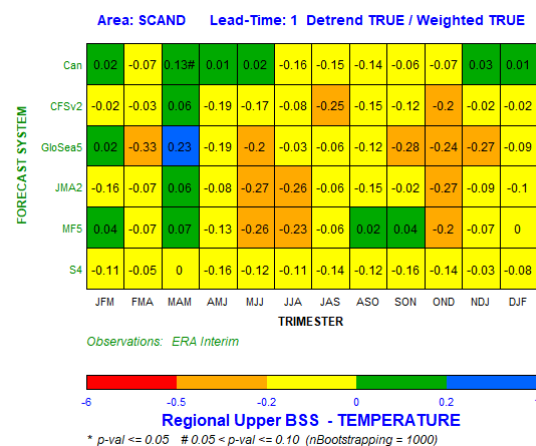
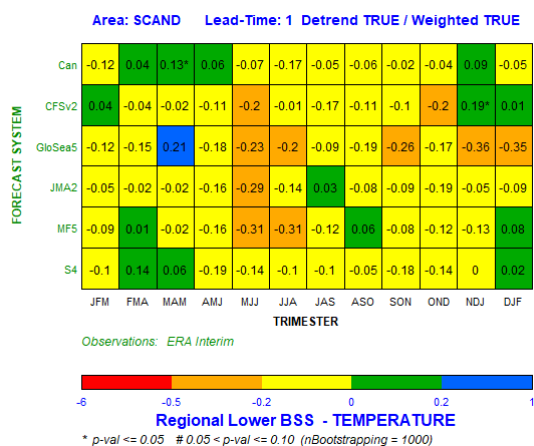
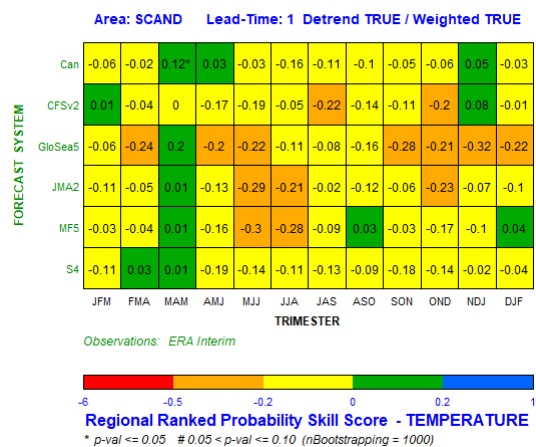
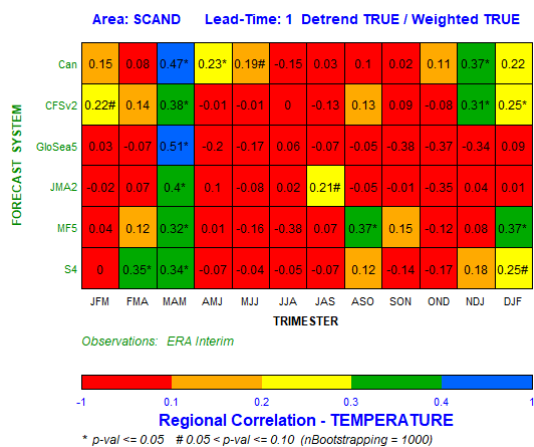


Table 86. The same as Table 74, but for SCANDINAVIA domain

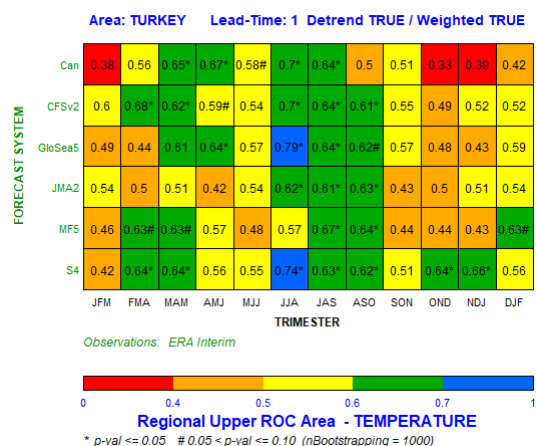
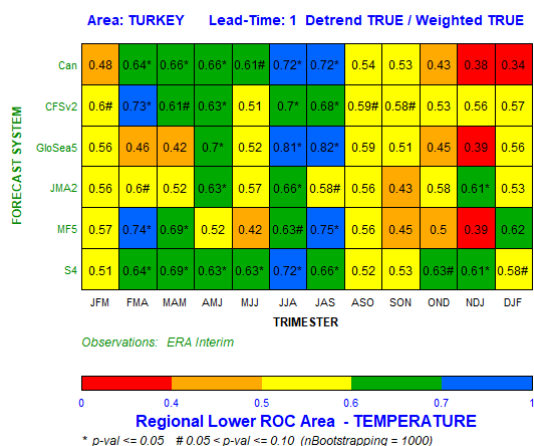
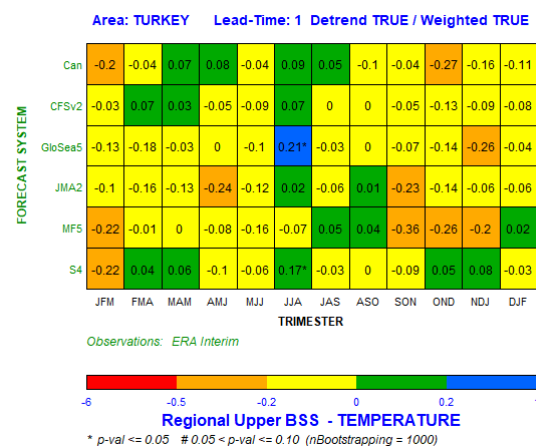
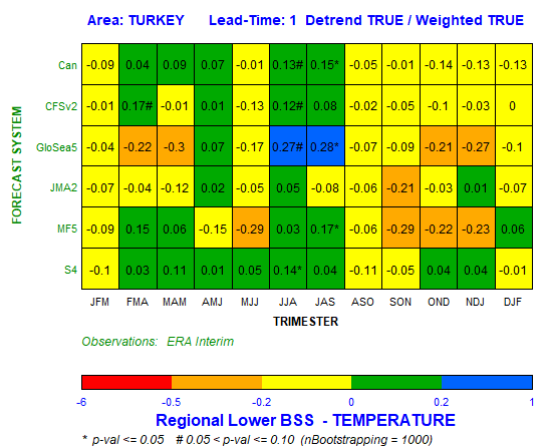
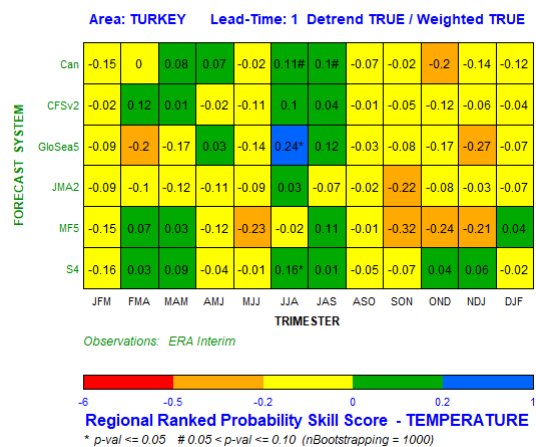
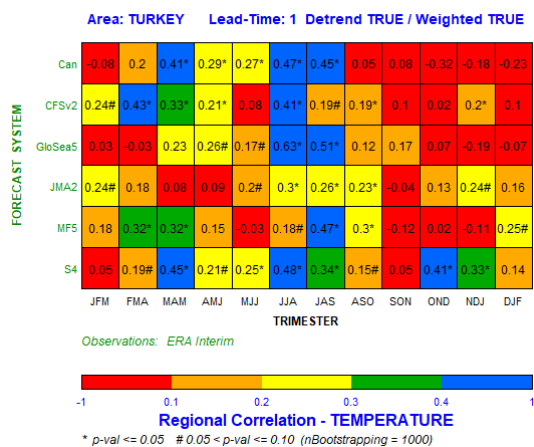


Table 87. The same as Table 74, but for TURKEY domain