Climate Dynamics (2006)
DOI 10.1007/s00382-006-0134-8

Jean-Philippe Boulanger · Fernando Martinez
Enrique C. Segura

Projection of future climate change conditions using IPCC simulations, neural networks and Bayesian statistics. Part 1: Temperature mean state and seasonal cycle in South America

Received: 2 June 2005 / Accepted: 20 February 2006

Abstract Projections for South America of future climate change conditions in mean state and seasonal cycle for temperature during the twenty-first century are discussed. Our analysis includes one simulation of seven Atmospheric-Ocean Global Circulation Models, which participated in the Intergovernmental Panel on Climate Change Project and provided at least one simulation for the twentieth century (20c3m) and one simulation for each of three Special Report on Emissions Scenarios (SRES) A2, A1B, and B1. We developed a statistical method based on neural networks and Bayesian statistics to evaluate the models' skills in simulating late twentieth century temperature over continental areas. Some criteria [model weight indices (MWIs)] are computed allowing comparing over such large regions how each model captures the temperature large scale structures and contributes to the multi-model combination. As the study demonstrates, the use of neural networks, optimized by Bayesian statistics, leads to two major results. First, the MWIs can be interpreted as optimal weights for a linear combination of the climate models. Second, the comparison between the neural network projection of twenty-first century conditions and a linear combination of such conditions allows the identification of the regions, which will most probably change, according to model biases and model ensemble variance. Model simulations in the southern tip of South America and along the Chilean and Peruvian coasts or in the northern coasts of South America (Venezuela, Guiana) are particularly poor. Overall, our results present an upper bound of potential temperature warming for each scenario. Spatially, in SRES A2, our major findings are that Tropical South America could warm up by about 4°C, while southern South America (SSA) would also undergo a near 2–3°C average warming. Interestingly, this annual mean temperature trend is modulated by the seasonal cycle in a contrasted way according to the regions. In SSA, the amplitude of the seasonal cycle tends to increase, while in northern South America, the amplitude of the seasonal cycle would be reduced leading to much milder winters. We show that all the scenarios have similar patterns and only differ in amplitude. SRES A1B differ from SRES A2 mainly for the late twenty-first century, reaching more or less an 80–90% amplitude compared to SRES A2. SRES B1, however, diverges from the other scenarios as soon as 2025. For the late twenty-first century, SRES B1 displays amplitudes, which are about half those of SRES A2.

J.-P. Boulanger (✉)
Tour 45-55/Etage 4/Case 100 UPMC,
LODYC, UMR CNRS/IRD/UPMC, 4 Place Jussieu,
75252 Paris Cedex 05, France
E-mail: jpb@lodyc.jussieu.fr
Tel.: +33-1-44275157

F. Martinez · E. C. Segura · J.-P. Boulanger
Departamento de Ciencias de la Atmosfera y los Oceanos
Facultad de Ciencias Exactas y Naturales,
University of Buenos, Aires, Argentina

1 Introduction

The Intergovernmental Panel on Climate Change (IPCC) published a Special Report on Emissions Scenarios (SRES) in 2000. This report describes the new set of emissions scenarios used in the Third Assessment Report. The SRES scenarios have been constructed to explore future developments in the global environment with special reference to the production of greenhouse gases and aerosol precursor emissions. While exhaustive description of the scenarios can be found elsewhere (Nakicenovic et al. 2000), it is worth recalling that they are based on a set of four narrative storylines labeled A1, A2, B1, and B2. The storylines combine two sets of divergent tendencies: one set varying its emphasis between strong economic development and strong environmental protection, the other set between increasing

globalization and increasing regionalization. Analyses of such scenarios in different regions of the globe performed for the Third Assessment Report have shown that, for a given model, the large scale geographical pattern of the simulated response to various forcing scenarios proved to be very similar, as only the amplitude of the response varied (Ruosteenoja et al. 2003). While in its Third Assessment Report, the IPCC concluded that the global average surface air temperature has increased by 0.6 ± 0.2°C during the twentieth century, it is also projected to increase by 1.4–5.8°C between 1990 and 2100. However, the projections from one region to another may differ significantly.

A major challenge now for the scientific community is to take advantage of the large ensembles of multi-model simulations provided by IPCC in the Fourth Assessment Report. Moreover, the importance of estimating the most probable climate change conditions in the different regions of the globe requires developing new statistical techniques, which will optimally combine the multi-model simulations based on their skills in simulating present climate conditions. Works by Giorgi et al. (2001), Giorgi and Mearns (2002), and Tebaldi et al. (2005) are based on Bayesian statistics and offer an interesting methodology to optimally combine models. A crucial step in the Bayesian approach is to choose correctly the prior distributions of the quantities of interest (Wigley and Raper 2001; Reilly et al. 2001; Allen et al. 2001; Forest et al. 2002). Considering the errors of CGCMs in simulating present-day climate, the most usual assumption is to give more importance in a multi-model approach to models with skill in simulating present climate conditions. However, the reader must keep in mind that a model may simulate well present-day climate and poorly respond to greenhouse gas forcing. Such a possibility is a caveat of the method.

In the present study, we aim at evaluating the evolution of large scale patterns rather than regionally averaged indices. Our strategy is therefore based on combining spatial maps of a multi-model ensemble. We decided to define a method based on the ability of each model to represent the large scale continental temperature. A solution to the present problem is the use of a neural network, whose parameters are optimized by Bayesian statistics (see Appendix). The neural network approach should lead (1) to determining which model contributes most to the output and (2) to extrapolating the optimal combination of models to twenty-first century conditions:

(a) The importance of an input to a trained variable is actually measured by the magnitude of the weights fanning out from the input (see Figs. 19, 20). If the weights are small, the input contributes little; if the weights are large, the input contributes more. The remaining question is to know what "more" means. In fact, the magnitude is relative to the other input weights. Thus, an interesting index, the model weight index (MWI), is the scaled inverse variance of the weights fanning out from each input. As a consequence, the MWI is a measure of the relative importance of each input to the trained dataset. For this reason, it can be also used as a weight for a linear combination of all the models, which can be compared to the neural network output.

(b) The neural network parameters (like those of any statistical method), also called weights, are optimized, based on a training dataset. If the distribution of the dataset changes dramatically, the method (and its parameters) usually does not project, i.e., extrapolate well. The neural network approach is usually considered to have good skills when the input data belong to a distribution similar or close to the distribution of the training dataset. In the case of temperature climate, all models predict an increase of the mean temperature in South America of various °C by the end of the twenty-first century. Such a big change produces a significant shift in the distribution and can impede the neural network from properly projecting twenty-first century conditions. In a case like that, it is interesting to analyze the linear combination and projection of models based on the MWIs (if shown to have a certain skill), which, to a certain extent, is a simplified linear version of the neural network. The neural network optimized through Bayesian procedures then offers two alternatives in estimating future climate changes. One alternative is to use the network directly as an extrapolator if it is proven to have an extrapolation skill. A second alternative is to combine the IPCC models linearly, using the MWIs.

Our work mainly focuses on South America as a contribution to the CLARIS European Project (http://www.claris-eu.org), but our method is universal and will soon be applied in different regions of the world.

Data, models, and scenarios used in the present study are described in Sect. 2. In Sect. 3, we discuss the method and introduce the MWI, which describes the weights of each model in the model combination. In Sect. 4, we focus on the seasonal temperature cycle projections. In Sect. 5, we present the projection of temperature mean states. In Sect. 6, we conclude and discuss the results and summarize the regional impacts of mean state and seasonal cycle changes.

2 Data, models, and scenarios

2.1 Data

The CRU TS 2.0 dataset comprises 1,200 monthly grids of observed climate and covering the global land surface at 0.5° resolution. There are five climatic variables available: cloud cover, DTR, precipitation, temperature, and vapor pressure. The temperature and precipitation data sets used are the 0.5° latitude/longitude dataset of monthly surface climate extending from 1901 to 2002

over global land areas, excluding Antarctica (http://www.cru.uea.ac.uk/~timm/grid/CRU_TS_2_0.html). The authors have already used a previous version (New et al. 2000) of this dataset (Boulanger et al. 2005), and have shown that, at least for precipitation, the comparison with satellite-based rainfall in South America was relatively good. Considering that we are mainly interested by large-scale patterns, the data are interpolated onto a 2.5×2.5° grid. Although this data set may present some differences to the Jones and Moberg (2003) data sets since urban effects have not been corrected in CRU TS 2.0, our spatial average on a 2.5×2.5° grid filters out a large part of this effect and does not affect the large scale pattern structures under study.

2.2 Models

We focused our effort on a multi-model analysis, considering only one simulation for each Atmospheric-Ocean Global Circulation Model (AOGCM). Moreover, we only considered models, which provided monthly precipitation and temperature outputs for twentieth century (20c3m), A2, A1B, and B1 scenarios. Overall, and although more IPCC models are certainly available now, our analysis was limited to a list of seven AOGCMs presented in Table 1. All the model outputs are interpolated over the 2.5×2.5° grid defined for the observations. Some models have finer resolutions, other have coarser resolutions, but overall the 2.5° resolution grid is a good compromise, which does not affect the large-scale patterns, and which allows a reasonable level of regional description.

2.3 Scenarios

The SRES scenarios are reference scenarios for the twenty-first century that seek specifically to exclude the effects of climate change and climate policies on society and the economy (“non-intervention”). They are based on a set of four narrative storylines labeled A1, A2, B1, and B2. The storylines combine two sets of divergent tendencies: one set varying its emphasis between strong economic development and strong environmental protection, the other set between increasing globalization and increasing regionalization (Nakicenovic et al. 2000). Our analysis made use of only three families briefly described as follows:

- A1: A future world of very rapid economic growth, low population growth, and rapid introduction of new and more efficient technology. Major underlying themes are economic and cultural convergence and capacity building, with a substantial reduction in regional differences in per capita income. In this world, people pursue personal wealth rather than environmental quality.
- A2: A differentiated world. The underlying theme is that of strengthening regional cultural identities, with an emphasis on family values and local traditions, high population growth, and less concern for rapid economic development.
- B1: A convergent world with rapid change in economic structures, “dematerialization” and introduction of clean technologies. The emphasis is on global solutions to achieving environmental and social sustainability, including concerted efforts for rapid technology development, dematerialization of the economy, and improving equity.

The storylines were quantified to provide families of scenarios for each storyline. In all 40 scenarios were quantified, six of which are used as illustrative scenarios by the IPCC, and we only considered three of them: A1B (balanced across energy sources), A2 (with a high-order radiative forcing), and B1 (more moderate radiative forcing).

Table 1 List of Atmospheric-Ocean Global Circulation Models

Model name and institute	Ocean model	Atmosphere model	Land model	Ice model	References
ipsl_cm4 *IPSL*	OPA8.1 2×2L31	LMDZ.3-96×72×19	ORCHIDEE1.3	LIM	
cnrm_cm3 *Météo-France*	OPA8.1 2×2L31	Arpege-Climat v3 (T42L45, cy 22b+)	TRIP	Gelato 3.10	Salas-Melia et al. (2004)
mpi_echam5 *MPI*	(1×1L41)	ECHAM5 (T63L32)		ECHAM5	Roeckner et al. (2003) Marsland et al. (2003) Haak et al. (2003)
ukmo_hadcm3 *UKMO*	1.25×1.25	2.5×3.75	MOSES1		Gordon et al. (2000) Johns et al. (1997)
ncar_ccsm3_0 NCAR	POP1.4.3, g×1v3	CAM3.0, T85L26	CLM3.0, g×1v3	CSIM5.0, T85	Collins et al. (2005)
gfdl_cm2_1 *GFDL*	OM3.1 (mom4p1p7_om3p5, tripolar360×200L50)	AM2.1 (am2p13fv, M45L24)	LM2	SIS	Delworth et al. (2005) Gnanadesikan et al. (2005) Wittenberg et al. (2005) Stouffer et al. (2005)
miroc3_2_medres *MIROC*	COCO3.3 256×192 L44	AGCM5.7b, T42 L20	MATSIRO T42	COCO3.3, 256×192 L44	

3 The neural network approach

3.1 General concepts

A detailed description of the multi-layer perceptron (MLP) and the procedures used in the present study are presented in Appendix. A brief summary of the method follows. In the present study, we will only focus on a two-layer network architecture (Fig. 19).

3.2 Training of the MLP architecture

The objective is to compare spatial temperature maps simulated by a set of climate models to observations. Therefore, we define:

- In the input layer: one neuron for the longitude grid point, one neuron for the latitude grid point and as many additional neurons as models (in the present case 7). Giving two neurons to the spatial location of each grid point makes it possible to take into account the model and data spatial dependence and correlation.
- In the output layer: one neuron for observations.
- In the hidden layer: a number of neurons to optimize.

The evidence procedure (Bayesian method) is used to train the MLP (see Appendix). Two methods (Bayesian approach and classical approach) are used to select the MLP architecture.

3.3 Multi-layer perceptron selection procedure

When the relationship between inputs and outputs is complex or when the dimension of the inputs is high, we found that the MLP optimization procedure may actually converge toward a local minimum rather than an "expected" absolute minimum. To avoid this problem, we decided to run the optimization procedure up to 50 times (we found this number to be large enough to always converge to a consistent result) for any given number of neurons in the hidden layer. Two cases among the 50 trials only differ by their randomly initiated weights. In these 50 cases, we always optimized the weights and hyperparameters using the evidence procedure. However, for each given number of neurons in the hidden layer, we selected two networks: (1) the one with the minimum negative log evidence value, called evidence index from now on (Bayesian approach) and (2) the one with the minimum error during a test period (classical approach). In the classical approach, a test period (here 1951–1975) different from the training period is selected and the MLP error after convergence is evaluated. The MLP with the minimum test error is then selected for any given architecture.

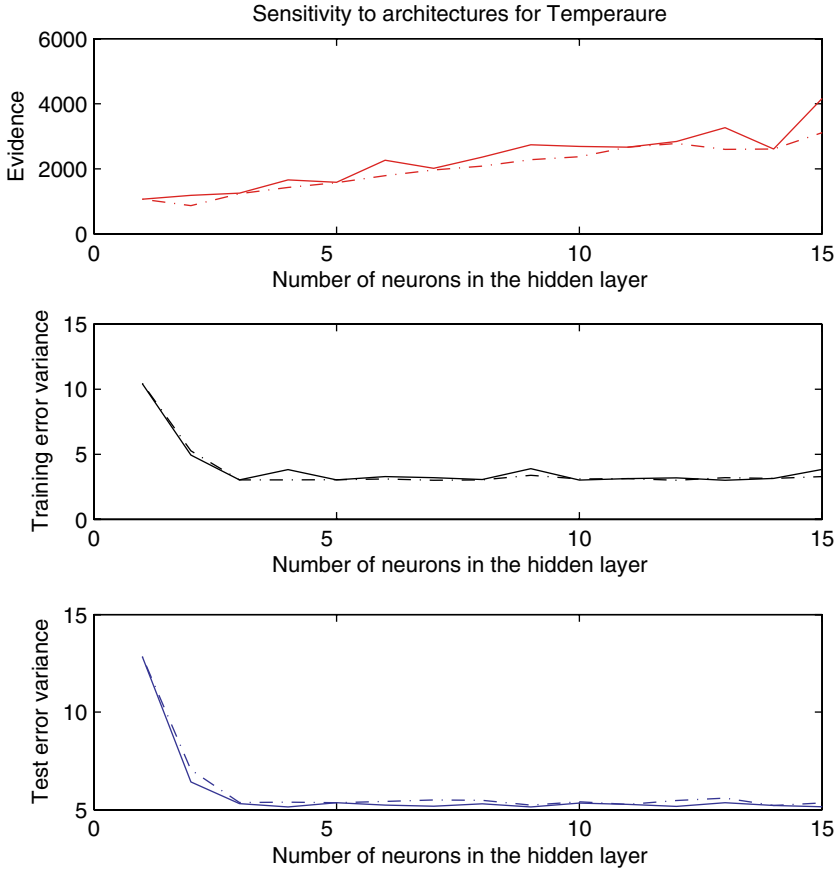


Fig. 1 Sensitivity of the evidence (*upper panel*), training error variance (*middle panel*, in $°C^2$) and test error variance (*lower panel*, in $°C^2$) to the number of neurons in the hidden layer. The training error variance is computed over the training period (1976–2000). The test error variance is computed over the 1951–1975 period. *Solid lines* represent the values for the networks selected using the classical approach. The *dashed lines* represent the values for the networks selected using the Bayesian approach

In general, the networks selected either by the Bayesian or classical approaches give similar results (Fig. 1). In the present case, we found the evidence index always increases with the architecture, in agreement with the fact that the increased number of neurons in the hidden layers penalizes the architecture. Moreover, we found the evidence index to be relatively similar, whichever of the two selected networks we study.

Figure 1 shows the training and test errors based on the mean annual temperature fields. The sensitivity to the architecture is similar whether we study the seasonal temperature fields or the mean annual field. In the case in Fig. 1, both the training and test errors decrease from 1 to 3 neurons and then remain relatively stable as a consequence of using the evidence procedure in the MLP parameter optimization, which reduces the over fitting risk. As our results (twenty-first century temperature projections) were not found to be sensitive even if a maximum of 10 or 15 neurons was considered, only the architectures between 3 and 10 neurons selected by the Bayesian or classical approach are taken into account in the present study (this represents an ensemble of 16 networks, i.e., projections). Finally, each architecture was weighted according to the inverse of its error during the test period as follows: (a) first, the test error variance of the ensemble was linearly scaled between 0 and 1; (b) second, the values are normalized. The weights are applied to the projection as well as to the MWIs in any mean or standard deviation calculation.

3.4 Method for twenty-first century projection

As stated in the introduction, the use of a neural network approach to combine climate models actually offers two alternatives to project twenty-first century climate conditions. One alternative is to use the network directly as an extrapolator if it is proven to have an extrapolation skill. A second alternative is to combine the IPCC models linearly with the MWIs.

In the ideal case (first alternative), the MLP could be used as an extrapolator. Unfortunately, in most cases, the non-linear nature of the network makes it impossible to use it as such. Indeed, extrapolation is much more reliable in linear models than in nonlinear models. In the present case, the major problem in using the MLP for extrapolation is that all IPCC models simulate a strong increase in continental temperatures at the end of twenty-first century, so the values of the input space under twenty-first century climate conditions do not appear in the training values meaning that the network has never learned such values. Therefore, the network may underestimate or poorly reproduce the potential increase of temperature. This hypothesis will be tested in the next section.

As to the second alternative, two questions naturally arise. (1) Why would the indices associated to a non-linear network optimization have skill to combine models linearly?. (2) In what way is the result dependent on the models chosen in the combination? These two questions cannot be answered a priori. The skill of the linear combination can only be evaluated a posteriori. As for any multi-model ensemble combination, our model combination is, by definition, dependent on the skills of the models taken into account. It is obvious that, if a model were found to be better than all others in simulating certain aspects of the climate variability, not taking it into account would certainly affect the twenty-first century climate projection negatively. It is therefore better to work with an ensemble of models as large as possible. As stated earlier, the present work aims at demonstrating the feasibility of using a neural network approach with Bayesian statistics to combine IPCC models. We plan in extending the present work to as many models as will finally be available in the IPCC data server.

4 Calibration to twentieth century observations

4.1 Temperature mean state

4.1.1 Model-data comparison

Figure 2 compares the temperature observations to the seven models. The models are mostly different with observations in the Amazonian basin and along the Brazilian coasts. It is likely that the different soil models used by the different AOGCMs explain the large differences of their response in the Amazonian basin. Other differences are also observed between 15 and 30°S west of the Andes. Some differences in the way the different models represent the southward extension of warm temperature east of the Andes and the meridional and zonal gradients are also worth noting.

4.1.2 Temperature MWI

As previously described, MLPs with three to ten neurons in the hidden layer were calibrated by comparing the temperature mean state over the 1976–2000 period simulated by the seven models to the one observed (maps displayed in Fig. 2). Figure 3 presents the temperature MWI as well as its uncertainty. It can be shown that, according to that index, the models, which contribute most to the output, are (in decreasing order) the Meteo-France (cnrm_cm3) model, the MPI (mpi_echam5) and the IPSL (ipsl_cm4) model. It is likely that the major weight found for the Meteo-France model is due to its relatively good amplitude over most of the Amazonian basin and the southern regions. The weak MIROC index may be associated to the large differences in the model amplitude and general patterns, while the weak GFDL index is more likely to be associated to the high warm temperatures over the Amazonian basin and a relatively large zonal gradient on the eastern part of the continent, differing from the observed weak zonal

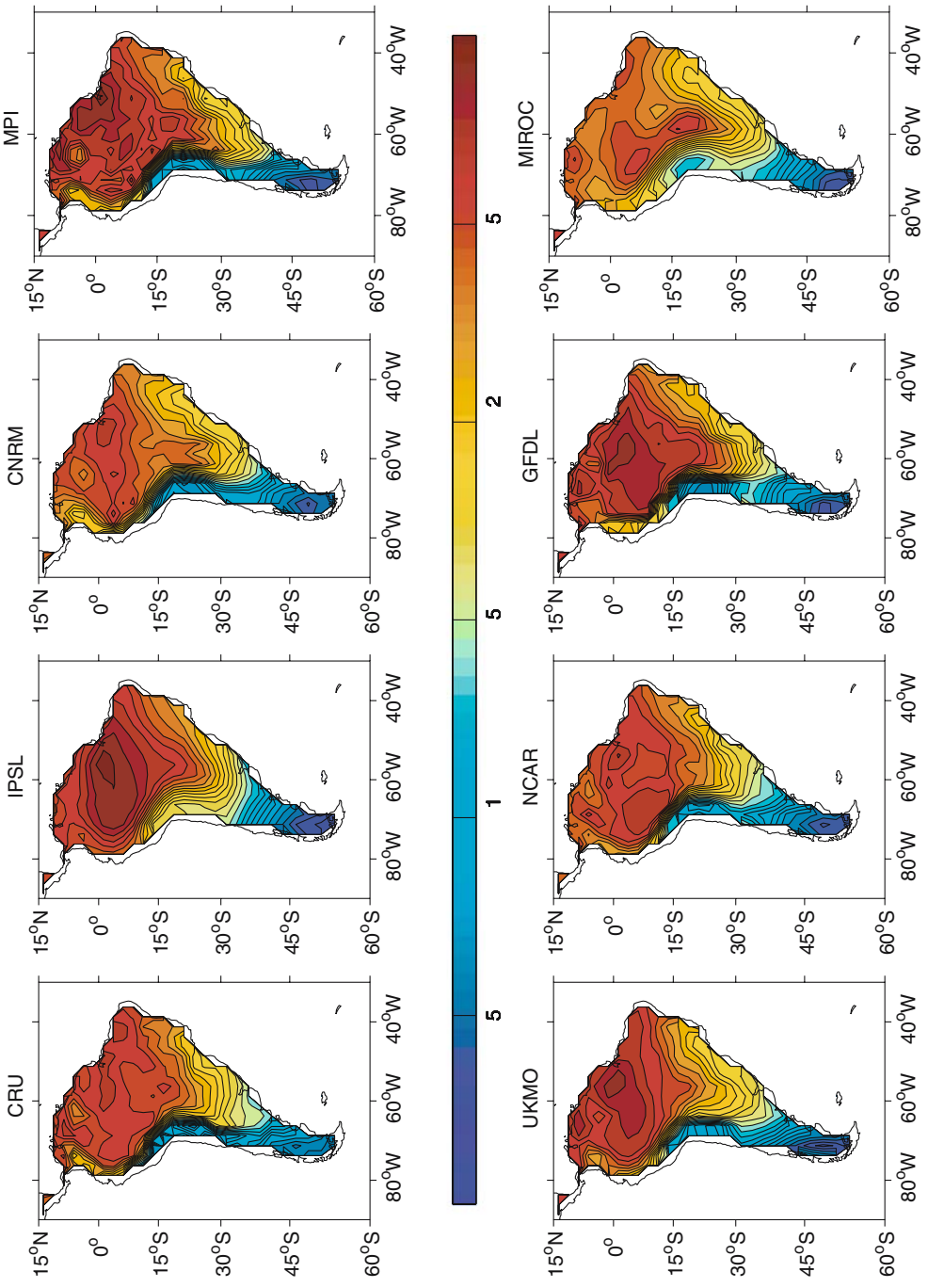


Fig. 2 Annual mean temperature for observations (CRU) and each of the seven models computed over the 1976–2000 period. Contours are every 1°C

gradient. The most striking result is the poor index of the UKMO model. Although it is difficult to identify exactly the reason for this, we suggest that it may result from: (1) a strong zonal gradient in temperature east of the Andes between 15 and 30°S where the observations have a rather weak zonal gradient; (2) cold temperatures in the southern tip of South America.

It is important to clarify that the index is not a quality index. It represents the model contribution in the mixing of all models when considering the entire continent. There is no doubt that more regional studies focusing on either the tropical or the subtropical region (such studies are beyond the scope of this paper) may give different results in the weighting of each model.

4.2 Temperature seasonal cycle

4.2.1 Model-data comparisons

Figures 4, 5, 6, and 7 show the shift of the four seasons (December–February, DJF; March–May, MAM; June–August, JJA; September–November, SON) from the mean state for CRU observations and the seven models.

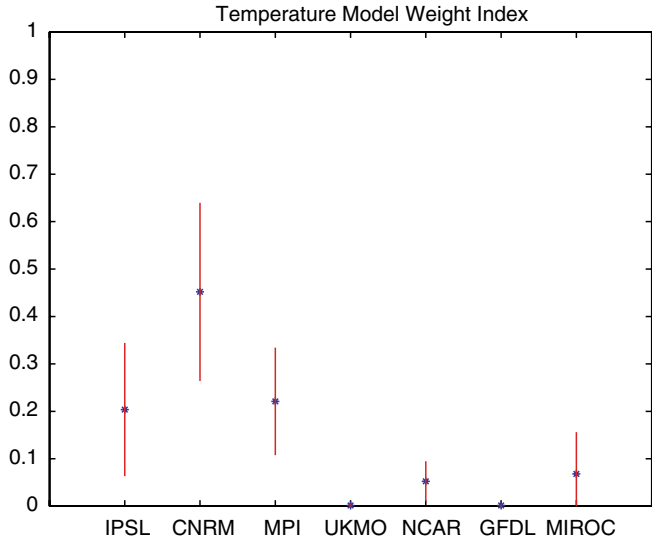


Fig. 3 Temperature model weight index computed for each model. Mean values are in *blue* and the error bar is in *red*

In austral summer (DJF; Fig. 4), the major differences between models and observations are as follows. First, the strong warm pattern centered near 35°S and between 70 and 65°W is often too strong (except for MIROC and NCAR) and usually located too far north (such as in UKMO). Second, most of the models overestimate the cold anomalies over the Amazonian basin (except CNRM, UKMO, and MIROC) or display large warm anomalies near the equator (IPSL, NCAR, GFDL).

In austral fall (MAM; Fig. 5), relatively weak anomalies are observed during that transition season. All the models tend to overestimate the anomalies in the Amazonian basin with some very large model-discrepancies in GFDL and NCAR. MIROC displays large cold anomalies near 30°S, while CNRM simulates cold anomalies in southern tip of South America.

In austral winter (JJA; Fig. 6), the observed pattern is similar to the DJF pattern but has the opposite sign. Some differences are observed in the absolute temperature amplitude as well as in the zonal gradient between the Andes and the Atlantic coast. Models also show an opposite pattern to DJF conditions although some models have more variability in their amplitude or gradients than the observations.

Finally, in austral spring (SON; Fig. 7), the observed and simulated patterns are not anti-symmetric to the MAM patterns. It appears that the models have a very strong bias in simulating the temperature anomalies over the Amazonian basin, and poorly simulate the warm anomalies extending along the Andes from 15°S to the southern tip of South America (except to a certain extent the MPI model).

4.2.2 Temperature MWI

As previously, the MLP was calibrated over the period 1975–1999. The temperature MWI is displayed for the four seasons in Fig. 8. As previously stated, this index measures the weight that the neural network attributes to that specific model in the function of transfer between IPCC models and observations. The major point observed in Fig. 8 is that, contrary to Fig. 4, error bars are relatively large meaning that from one specific architecture to another the weighting amplitude varies a lot so that it is much more difficult to identify the reasons why the method weights more or less one model or another. Anyway, we can draw some suggestions from Fig. 8 and Figs. 4, 5, 6, 7. First, the NCAR and GFDL model indices are significantly weak during most of the four seasons as compared to the model with the largest index. We tend to believe that this behavior is associated to their complex patterns over the Amazonian basin. The relatively better results of the GFDL model in JJA may actually be associated to a better simulation of the location and amplitude of the coldest anomalies in the subtropical region, which may have compensated the bias in the Amazonian basin. The relatively non-significant differences between the various model indices suggest that the method takes advantage of the different model patterns in its fit to the observations, although the twenty-first century projection may actually display larger ensemble errors (to be discussed later).

4.3 Neural network versus linear combination

Figure 9 shows the mean of our ensemble and the difference to observations, and the ensemble error (or standard deviation) both for the training and test periods. It shows that the MLP Ensemble corrects fairly well the different model biases in order to recover a large-scale pattern in good agreement with annual mean temperature observations. The differences with the observed mean state show a relatively patchy structure with values usually between ± 1°C. Moreover, all the MLPs are consistent in their reproduction of the 1976–2000 training period conditions displaying an ensemble error lower than 0.4°C. When the MLPs are used to project the test period (1951–1975), the pattern of differences is similar although slightly larger. It must be noted that the simulated differences between the two periods are very different from the observed differences. In order to test whether this result is an error of the method (overfitting) or an error of the models used as inputs, we computed the linear ensemble mean for the two periods with the MWIs. When representing the 1976–2000 period, interestingly the linear ensemble mean reproduces fairly well the observed features although it displays larger differences to observations than the MLP ensemble mean with structured patterns. These patterns represent model biases in the Amazons, over the Nordeste, in the La Plata Basin (LPB) as well as west of the Andes. Moreover, the linear ensemble mean error is much larger than the MLP consistent with the large differences between the models in certain regions (Fig. 2), and the MWIs display of large standard

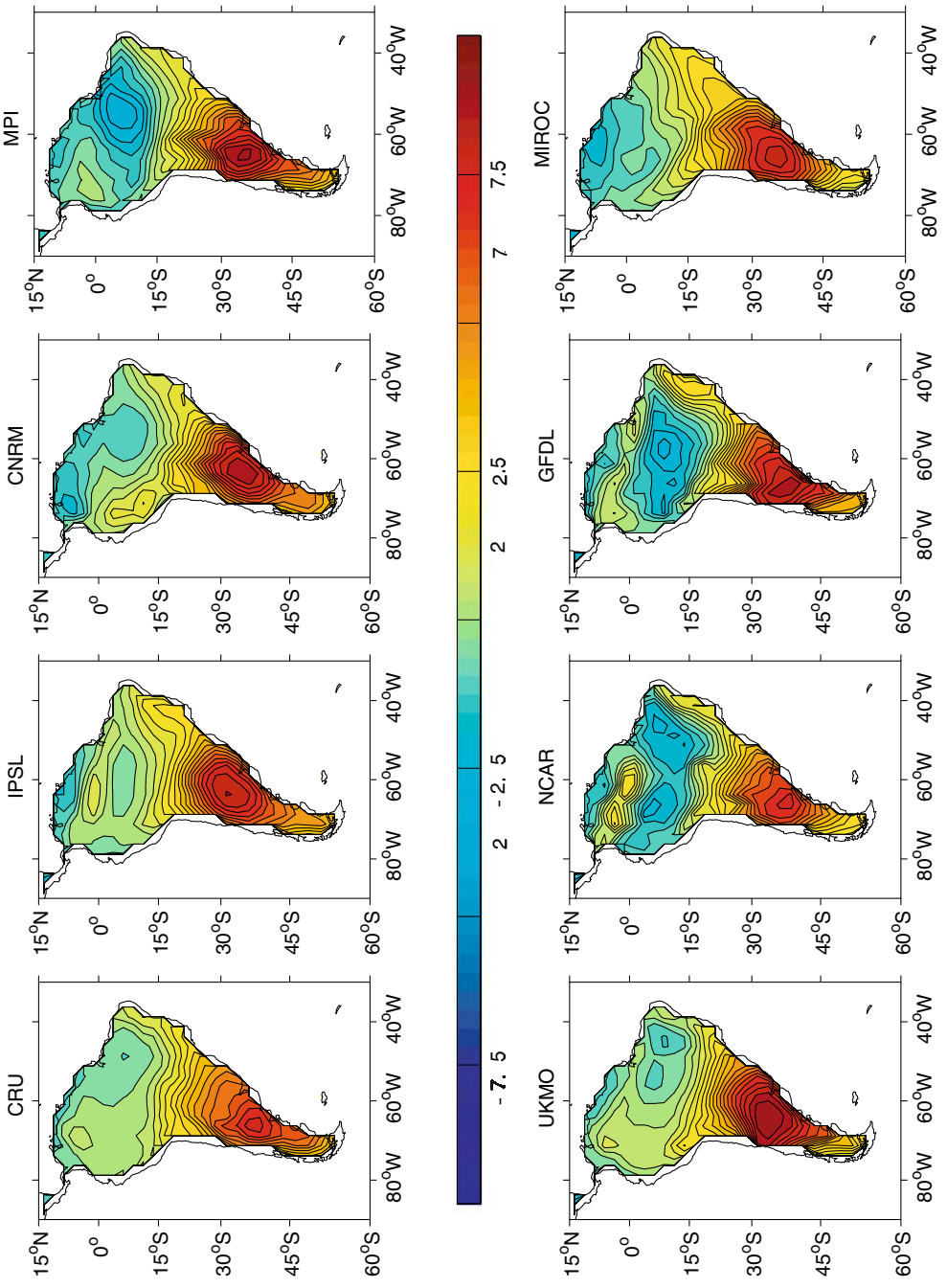


Fig. 4 Same as Fig. 2 but for the DJF season. Contours are every 0.5°C

deviations for certain models (Fig. 3). Interestingly, when the linear ensemble mean is computed over the 1951–1975 period, it is seen that the differences between the two periods are more similar than the ones displayed by the MLP, although of larger amplitude. However, the changes in the simulated patterns are very different than those observed. In fact, major changes in the precipitation regimes have been observed since the 1960s–1970s. Such changes are not simulated by the IPCC coupled GCMs. Moreover, the fact that both the MLP and linear ensemble means are similar shows that the MLPs are not overfitted.

Finally, we found the level of fit of the MLP for each of the four seasons during the training and test periods to be similar (not shown).

Before analyzing in detail the twenty-first century projections (Sect. 5), it must be seen whether the MLP has extrapolation skill. Figure 10 compares the respective behavior of the MLP and the linear ensemble, only considering as inputs the IPCC model mean temperature simulated under SRES A2 scenario for the 2076–2100 period. There are big differences between the two ensemble projections. The linear ensemble displays a strong warming with values ranging between 2 and 4°C,

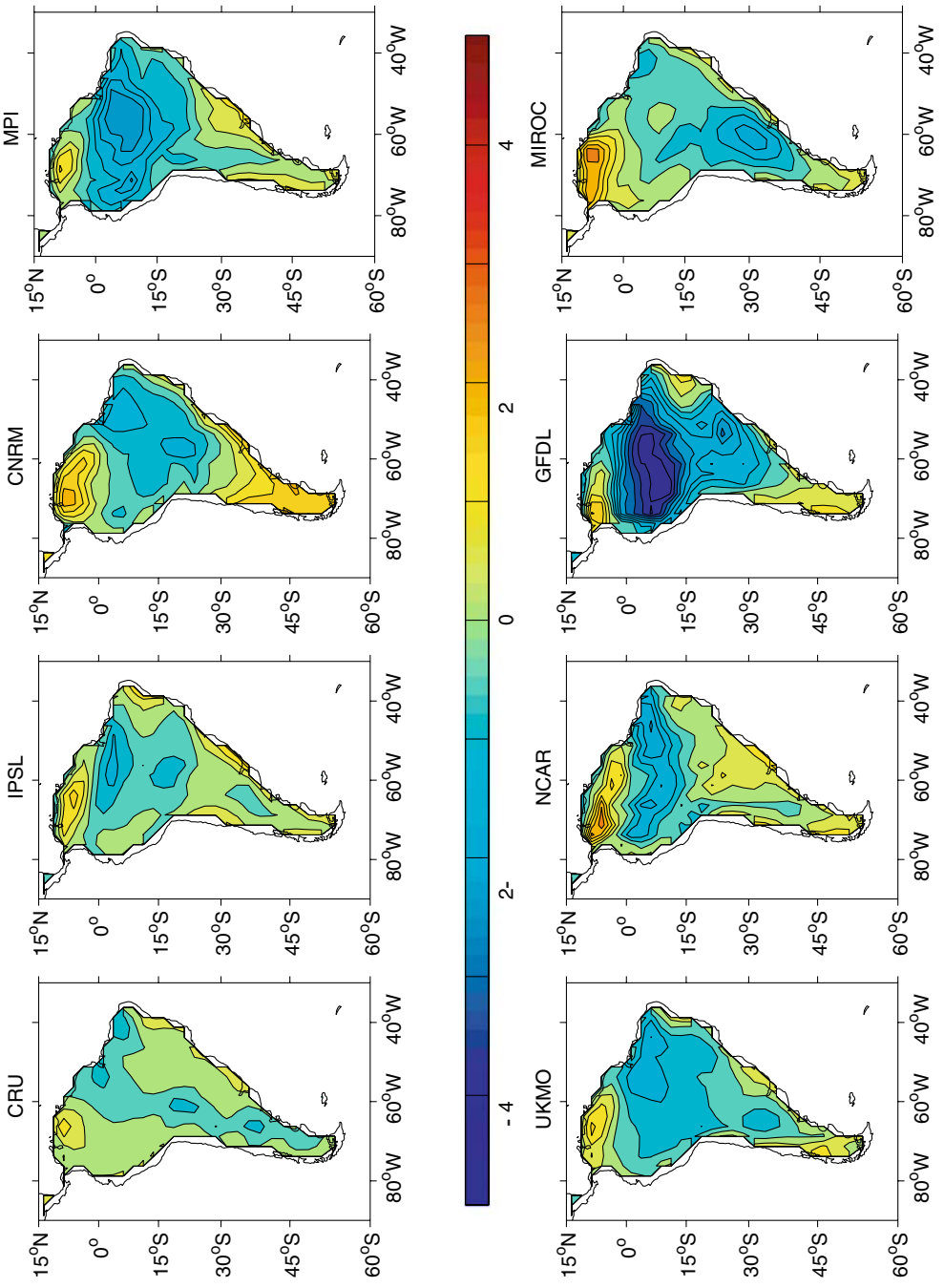


Fig. 5 Same as Fig. 4 but for the MAM season

while the MLP projection is much weaker and reaches a temperature increase of 4°C only, in specific regions (Pacific coast and parts of the Amazons and Southern Brazil). The agreement ratio between the MLP and the linear projections (Fig. 10,"Confidence Level") shows that the regions where the two methods converge (ratio more than 0.8) are the Pacific coast from Colombia to northern Chile (15°S), the Atlantic coast (from 15 to 35°S) with intermediate values (between 0.6 and 0.8) covering most of Brazil. In all other regions of South America, the ratio is low and can reach values close to 0.

In Tebaldi et al. (2005), the posterior distribution of the parameters used in the climate change linear projection is sensitive to two factors: the bias criterion (how the IPCC models represent present climate observations) and the convergence criterion (how the IPCC models agree in their climate change response). Considering that the MLP weights are built on present-day climate, we analyzed two criteria: the bias criterion (how the IPCC linear ensemble compares to 1976–2000 observations; Fig. 10) and the divergence criterion or inter-model variance (how the IPCC models differ from each other in simulating present-day climate; Fig. 10). The bias criterion presents significant errors west of the Andes from 10 to 35°S as well as in the southern tip of South America. In both cases, the confidence level is

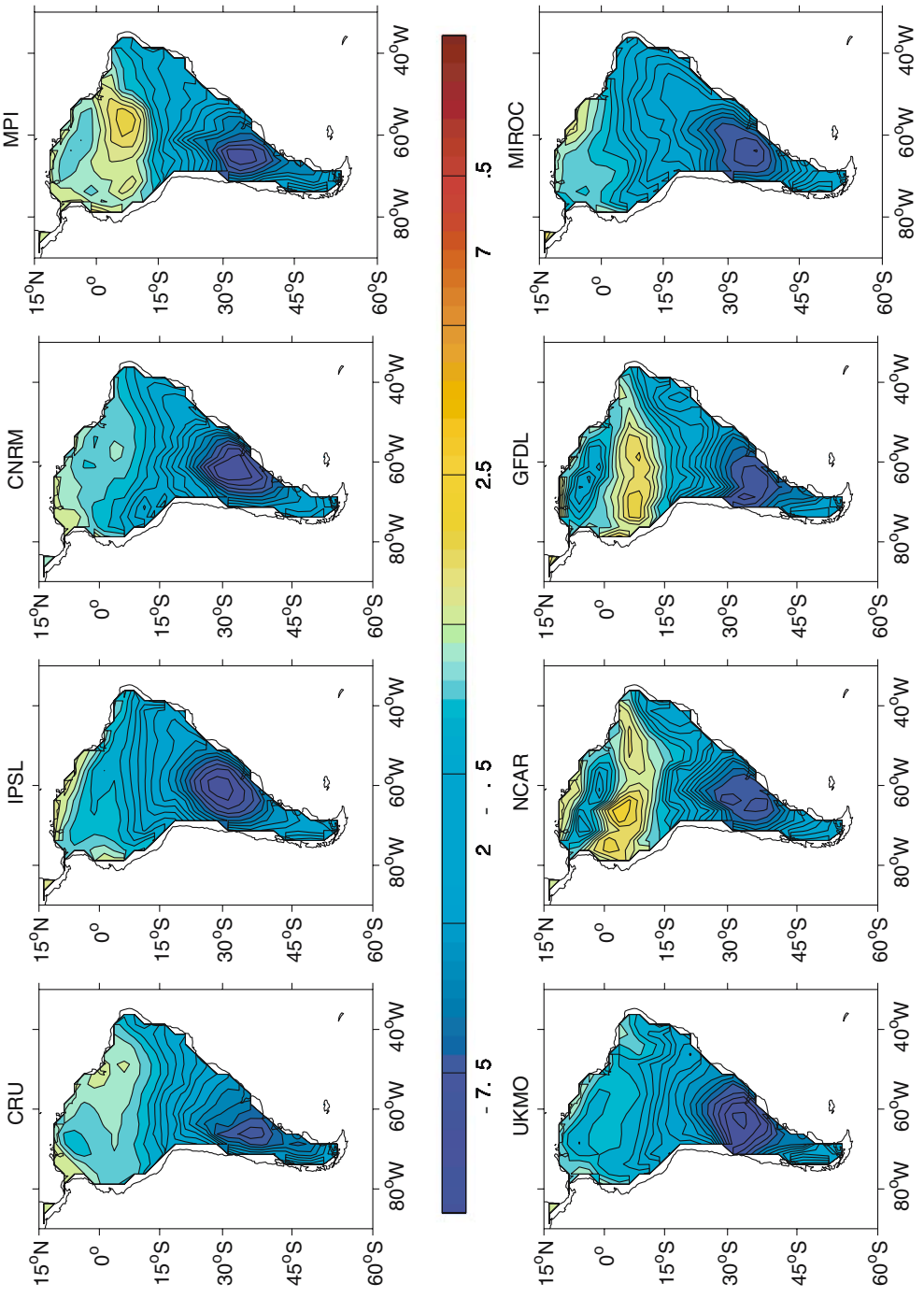


Fig. 6 Same as Fig. 4 but for the JJA season

close to zero. The divergence criterion displays significant errors along the Pacific coast as well as over part of the Amazons, in Guyana, Venezuela, and Colombia. In these regions, the confidence level is low or even close to zero displaying similar contour patterns.

In conclusion, and to a certain extent by analogy to the work by Coelho et al. (2004), the MLP penalizes the projected warming when either the linear combination error to observations or the IPCC inter-model variance are large. One could linearize the MLP projected change as follows: $\mathrm{MLP}(x,y) = P(x,y)\sum_{n=1}^{N}\mathrm{MWI}(n)\mathrm{IPCC}(n,x,y)$, where $\mathrm{IPCC}(n, x, y)$ represents the anomalous spatial map of the nth IPCC model considered in the study (difference between twentieth and twenty-first century conditions), $\mathrm{MWI}(n)$ is the model weight index associated to the nth IPCC model and $P(x, y)$ is a penalizing function, which could be written as: $P(x, y) = \exp(-\mathrm{Ve}(x, y)) \times \exp(-\mathrm{Vm}(x, y))$, where Ve and Vm are respectively the normalized variance of the linear ensemble error (bias criterion) and the normalized variance of the IPCC models (divergence criterion). Wherever Ve or Vm are different from 0, the MLP will penalize the twenty-first century projection. Such

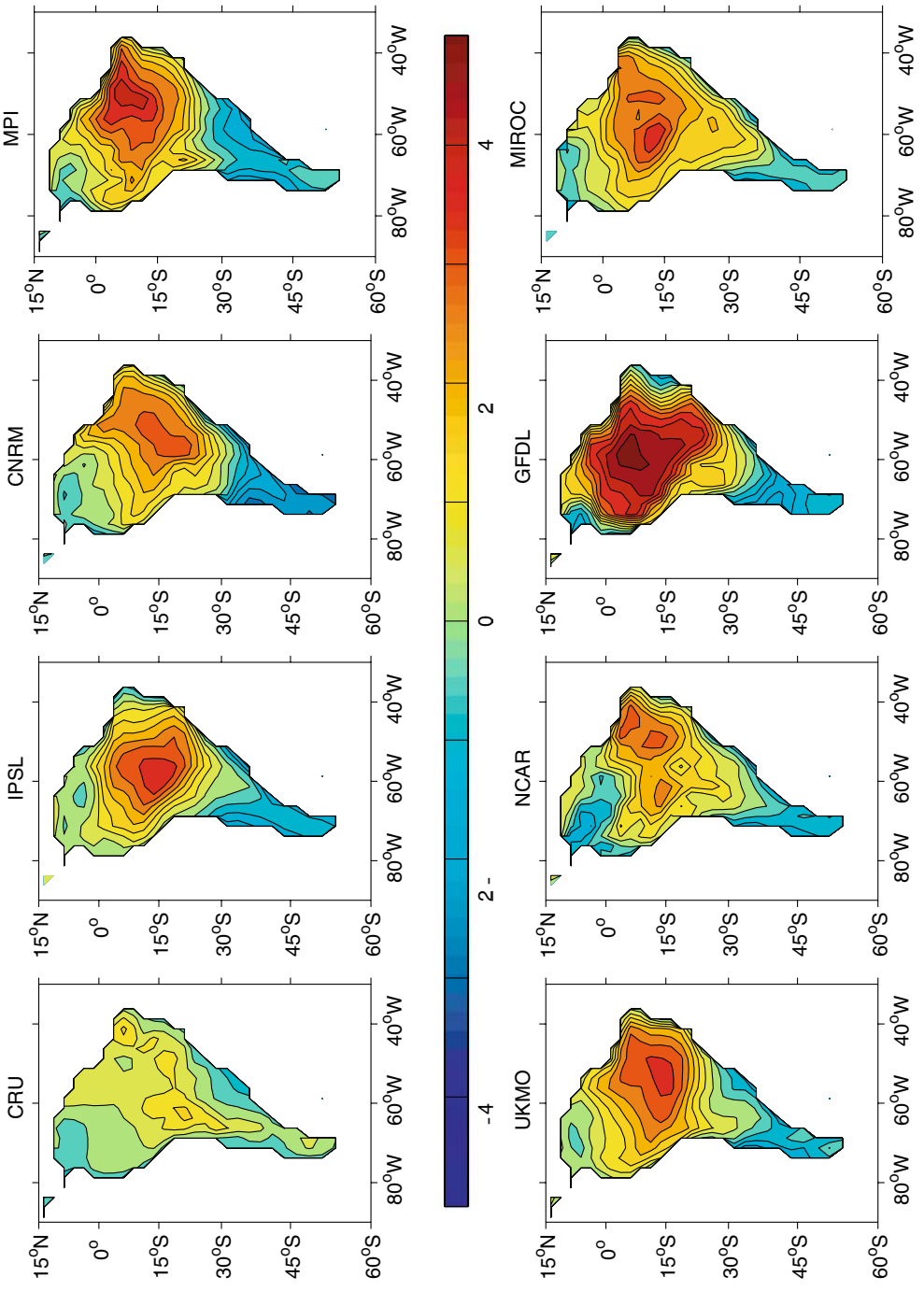


Fig. 7 Same as Fig. 4 but for the SON season

behavior is obvious in the Southern tip of South America, along the Chilean and Peruvian coasts as well as in the Guyana, Venezuela, and Colombia region (the shape of the weak warming there is strikingly similar to the shape of the divergence criterion).

Finally, the MLP projections underestimate the potential climate change projections simulated by the IPCC model, as the MLP penalizes the model projections according to the two kinds of errors described earlier. However, such behavior is also an advantage as by comparison to linear projection, it makes it possible to compute the spatially dependent confidence level of the changes in climate conditions (Fig. 10). Therefore, both the MLP and linear projections (based on MLP weights) should be analyzed jointly. The linear ensemble projection displays a possible climate change pattern, while the MLP helps to compute the confidence level of such a change.

Here below, we will only show the linear combination of IPCC models based on the MWIs. However, the linearly projected patterns are upper bounds of climate change amplitudes, and the confidence in these changes is limited to the regions where the confidence level in Fig. 10 is close to 1.

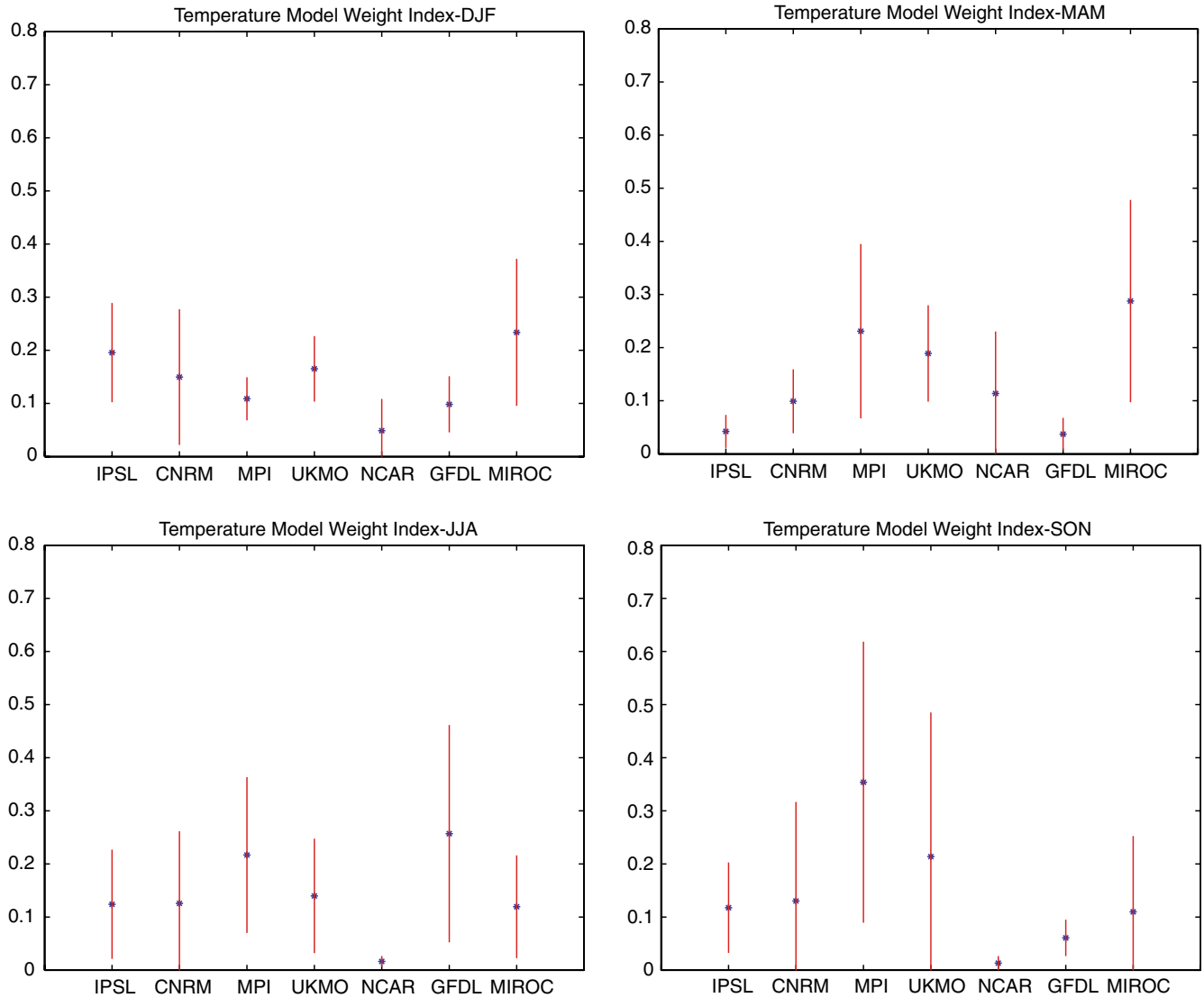


Fig. 8 Same as Fig. 3 but for the four seasons

Finally, in Fig. 11, we present the linear combinations of each season. It can be seen that the model linear combination presents various biases. In austral summer (DJF), the general structures are fairly well reproduced, but the amplitudes of the large temperature anomalies observed between 15 and 30°S are actually simulated too far to the north. Moreover, the decrease in temperature from the interior of the continent to the Atlantic Ocean north of 30°S is too strong in the ensemble, leading to a strong negative bias. Similar patterns are still present in austral fall (MAM) but with a reduced amplitude. In austral winter (JJA), the major bias is observed in the Nordeste and in the LPB with too cold an ensemble, and in the southern tip with too warm temperatures. Finally, in austral spring (SON), the model linear combination is way too warm over the Amazons and too cold in the southern tip of the continent. The ensemble error displays a pattern strongly related to the regions of larger differences. Overall, the ensemble error is in the order of or weaker than 1°C.

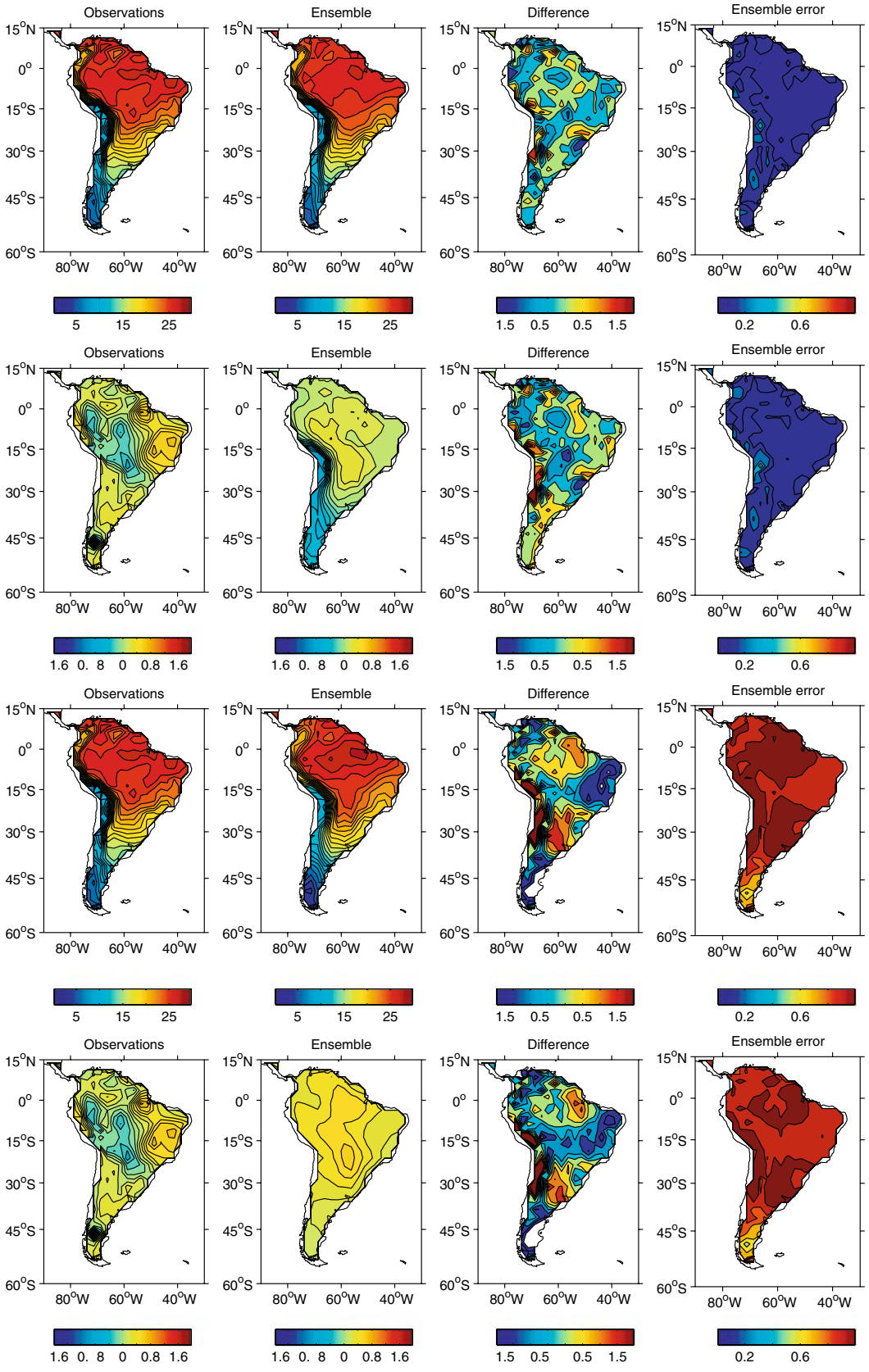
In conclusion, in decreasing order of magnitude, the most sensitive regions are the Amazon basin (with large anomalies except in winter), the LPB (during the extreme phases of the season), the southern tip of the continent and the Colombia–Venezuela–Guyana region.

5 Twenty-first century projection

5.1 Temperature mean state

5.1.1 Model-projection comparison

First, we compare the mean temperature projection given by the method when mixing the seven model outputs for the three scenarios A2, A1B, and B1 and for the four



25-year periods (2001–2025, 2026–2050, 2051–2076, and 2076–2100). For the simplicity's sake, we will only compare scenario A2 during 2076–2100 (Fig. 12). First, we can observe that all the models display a much warmer temperature mean state than during the 1976–2000 period over the Amazonian basin. This warming is

◄

Fig. 9 From *top to bottom* for each row of panels: (1) *From left to right* Observed 1976–2000 annual mean temperature (contours are every 1°C), neural network projection based on the 1976–2000 period (contours are every 1°C), differences between the projection and the observations (contours are every 0.5°C), ensemble variance (contours are every 0.1°C). (2) *From left to right* Observed differences between the 1976–2000 and 1951–1975 annual mean temperatures (contours are every 0.2°C), differences between the neural network projection based on the 1976–2000 period and the 1951–1975 period (contours are every 0.2°C), ensemble variance for the 1951–1975 projection (contours are every 0.1°C). (3) *From left to right* Observed 1976–2000 annual mean temperature (contours are every 1°C), linear projection based on the 1976–2000 period (contours are every 1°C), differences between the linear projection and the observations (contours are every 0.5°C), ensemble variance (contours are every 0.1°C). (4) *From left to right* Observed differences between the 1976–2000 and 1951–1975 annual mean temperatures (contours are every 0.2°C), differences between the linear projection based on the 1976–2000 period and the 1951–1975 period (contours are every 0.2°C), ensemble variance for the 1951–1975 linear projection (contours are every 0.1°C)

actually relatively general over most of the continent. We found that these patterns were mostly identical whatever the period under study as only the amplitude of the response varied. Moreover, when comparing the different scenarios (A2, A1B, and B1), we also found that the projected patterns were similar and differed mainly in their amplitude. While all the models suggest a warming of at least 2°C in the southern tip of the continent and of 3–4°C in the northern regions, large differences are observed between the models. NCAR displays the weakest warming amplitude, while UKMO exceeds the 5°C warming over tropical South America. Most of the models suggest a greater warming in the Colombia–Venezuela–Guyana region of 3–5°C. Then, depending on the model, strong warming patterns can be observed over the Amazons basin, the LPB, Nordeste or Chilean coasts. The ensemble projected pattern is now discussed.

5.1.2 Twenty-first century projection

Figure 13 displays the twenty-first century SRES A2 projected mean temperature for the four 25-year periods (2001–2025, 2026–2050, 2051–2076, and 2076–2100), the difference to the 1976–2000 pattern and the ensemble error. We can see clearly a warming over the continent after the 2026–2050 period. The major patterns associated to the warming are:

1. The tropical Pacific coastal warming associated to a warming in the IPCC models of the Pacific ocean coastal sea surface temperatures along the coasts of Equator, Peru, northern Chile, and Colombia, the warming reaches 4°C in certain regions;
2. A strong warming over southern Venezuela and northern Brazil;
3. On the eastern side of the Andes, a major warming (around 4°C), which covers a large part of the Amazons, and, which extends over most of the continent, slightly decreasing eastward and more strongly southward. Overall, the warming is higher than 2°C at each grid point.

However, as pointed out earlier, the actual amplitude of the future warming is very uncertain given the differences between the models (Fig. 12) and the confidence level displayed in Fig. 10. The main conclusions one can actually make, given such levels of uncertainty, are that the entire continent is likely to warm with a stronger amplitude in the tropical region than in the southern part, that the warming should be close to 3–4°C along the tropical Pacific coast and 2–3°C along the Atlantic coast where the confidence level is high (Fig. 10). In the Amazon basin, significant differences certainly result from land surface model physics.

The same projection for scenarios A1B and B1 for the last 25-year period is displayed in Fig. 14. We find the method to be relatively consistent as the warming patterns are very similar to the ones projected for scenario A2. The figures only differ in amplitude. SRES A1B warming amplitude pattern is intermediate between the 2051–2075 and 2076–2100 SRES A2 patterns. SRES B1 warming amplitude pattern is intermediate between the 2026–2050 and 2050–2076 SRES A2 patterns.

Overall, the projected continental mean warming is close to 4°C for SRES A2, while the same projections for the scenarios A1B and B1 are respectively 3.4 and 2.2°C. The ensemble error bar on the continentally averaged annual mean temperature rise is actually relatively small (on the order of 0.1°C), but it does not take into account the bias and divergence criteria, nor the discrepancies between model projections.

5.2 Temperature seasonal cycle

For the sake of clarity and brevity, we do not present here the comparison between the ensemble and the evolution of each model temperature patterns for each 25-year period and each of the four seasons but. concentrate on the description of the late twenty-first century projection.

5.2.1 SRES A2

Figure 15 shows late twenty-first century SRES A2 projection for each season. Some striking large-scale patterns can be observed:

1. First, in DJF, the largest warming is observed along the coasts of Chile and Peru (4°C) and over Venezuela and part of Colombia (4.5°C). Most of the Atlantic coast could experience a near 3°C warming as far south as 4°S. In the southern tip of the continent, the warming may reach from 1 to 2°C. Thus, the warming over northern South America (NSA) could get to 3.5°C (Fig. 16) and 3.7 and 3.1°C over southern South America (SSA) and the La Plata Basin (LPB).

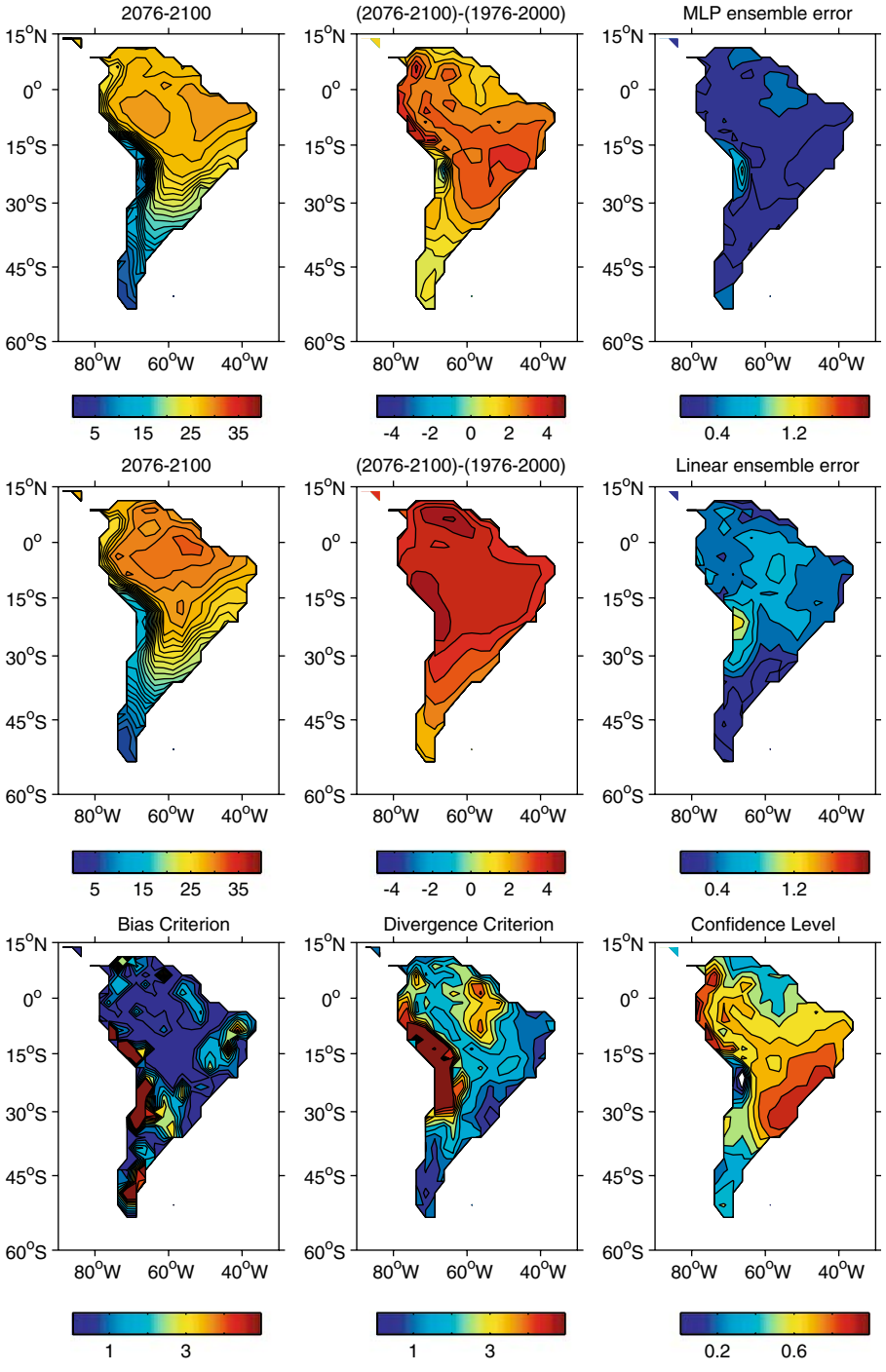


Fig. 10 From *top to bottom* for each row of panels: (1) *From left to right* MLP ensemble projection for late twenty-first century conditions (SRES A2; contours are every 1°C), differences with present-day climate conditions (contours are every 0.5°C) and ensemble variance of the projection (contours are every 0.1°C). (2) *From left to right* Linear ensemble projection for late twenty-first century conditions (SRES A2; contours are every 1°C), differences with present-day climate conditions (contours are every 0.5°C) and ensemble variance of the projection (contours are every 0.1°C). (3) *From left to right* Spatial representation of the bias criterion (squared error between observations and the linear ensemble computed for twentieth century conditions; amplitudes are in $°C^2$), of the divergence criterion (IPCC model dispersion or variance; amplitudes are in $°C^2$), of the confidence level (ratio between the MLP and linear projections; values are between 0 and 1)

2. In MAM (Fig. 15), the major temperature increase would be the same as seen over Colombia and Venezuela and the coasts of Chile and Peru. A warming of about 4°C could be located in southern Brazil. The meridional gradient of the warming trend may be weaker in MAM than in DJF. Thus, the southern tip of the continent could experience a warming of more than 2°C. In the three selected regions, the averaged warming could be 3°C in SSA, 3.9°C in NSA, and 3.7° in LPB.
3. In JJA (Fig. 15), the warming trend displays a strong meridional gradient near 30°S. North of 30°S, the warming could be strong with higher values over the Amazon and the northern coasts of South America. It could reach an average of about 4.5°C. South of 30°S, the average warming could be (it seems the place is missing here) around 2.6°C, while in LPB located part north and part south of 30°S, the warming could reach 3.6°C.
4. Finally in SON (Fig. 15), the highest warming (around 5.5°C) would be observed over the center of the Amazon and part of the Colombia–Venezuela region. The meridional gradient could be weaker than

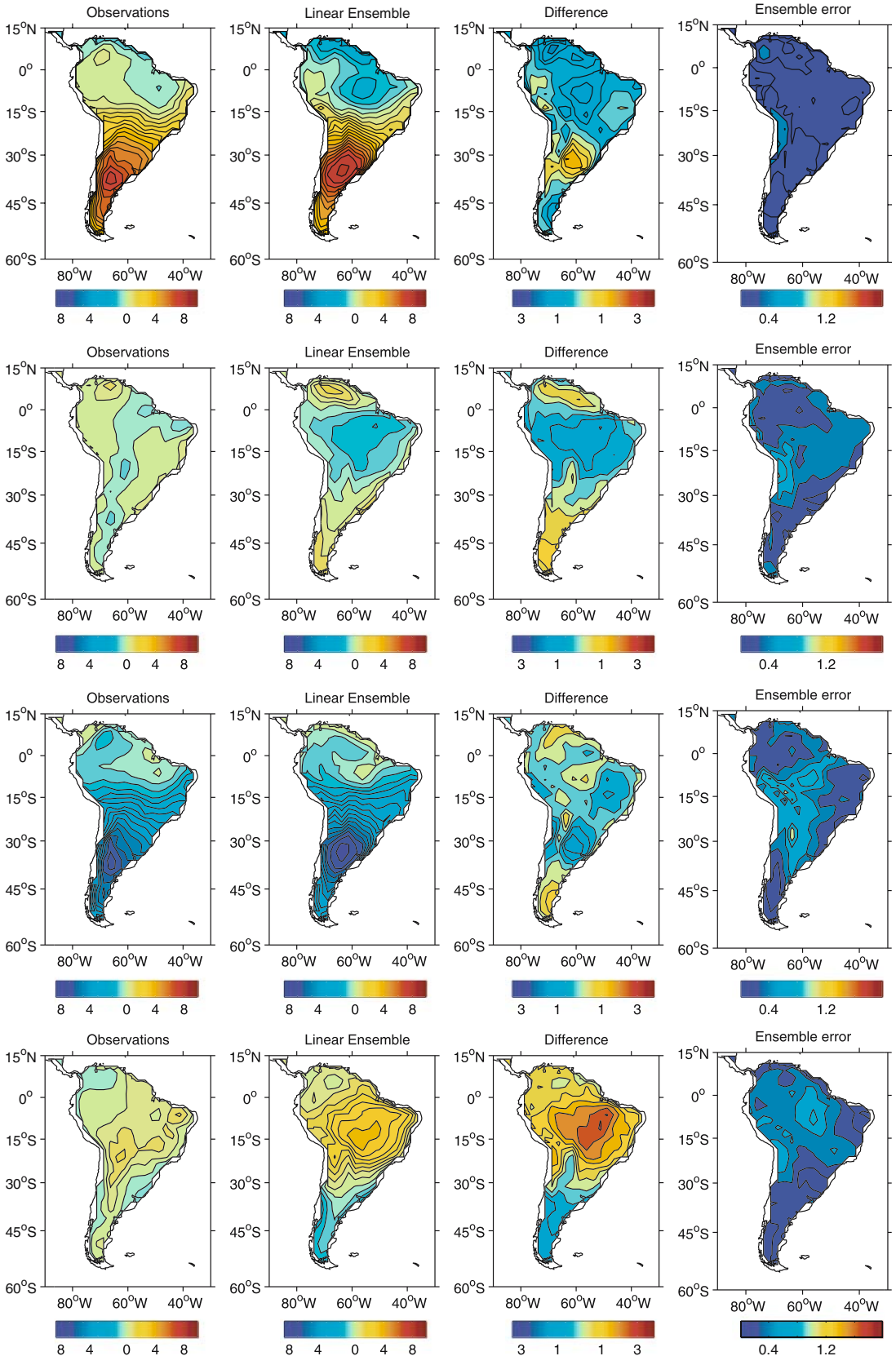


Fig. 11 From *top to bottom*, each row represents, for each season (DJF, MAM, JJA, SON), the observed seasonal anomalies computed over the 1976–2000 period (contours are every 1°C), the linear ensemble mean value (contours are every 1°C), the difference (contours are every 0.5°C), and the ensemble variance (contours are every 0.1°C)

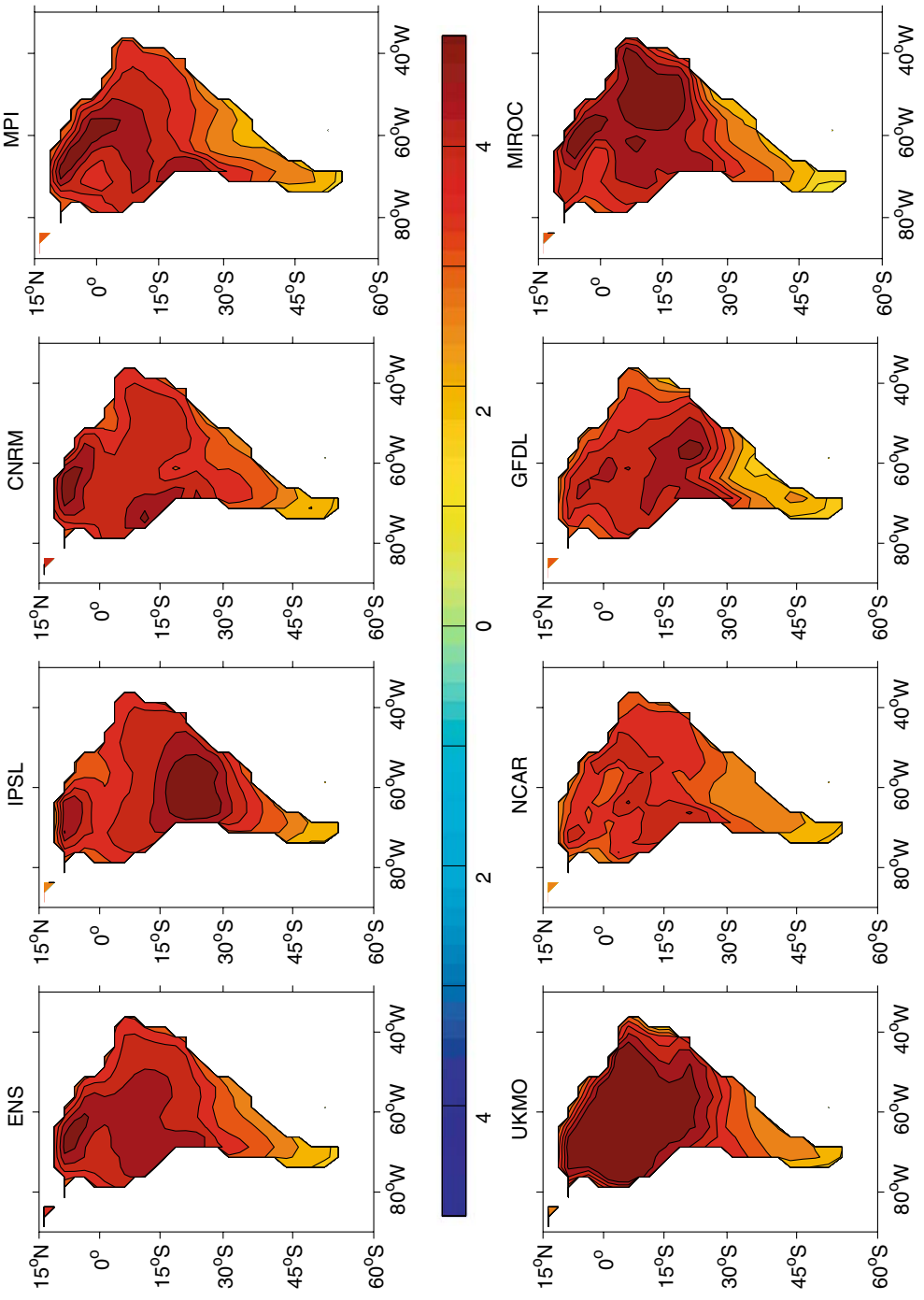


Fig. 12 2076–2100 SRES A2 annual mean temperature change projected by the linear method compared to the annual mean temperature change simulated by each model. Values are relative to the 1976–2000 period. Contours are every 0.5°C

in DJF, but the warming could remain much greater in NSA (4.7°C) than in SSA (2.9°C) or LPB (3.8°C).

Before concluding this analysis, it is interesting to highlight that the mean temperature-warming trend observed in each selected region (3.0°C in SSA, 4.3°C in NSA and 3.7°C in the LPB) is actually modulated along the course of the seasonal cycle. For instance, in SSA, the warming could be stronger in austral summer and fall (3.4 and 3.0°C) and weaker in winter and spring (2.6 and 2.9°C) suggesting greater amplitude of the seasonal cycle. In NSA, the warming could be weaker in austral summer and fall (4.1 and 3.9°C) and stronger in winter and spring (4.5 and 4.7°C) suggesting smaller amplitude of the seasonal cycle. Of course, the result is a very large-scale index and the modulation of the seasonal cycle may vary between sub regions. Finally, in the LPB, which is located on the edge of NSA and SSA, the trend is found to be relatively uniform over the seasons. It is worth pointing out that, at latitudes bordering between the tropical and subtropical climates, the projected large warming and especially a much warmer winter may have

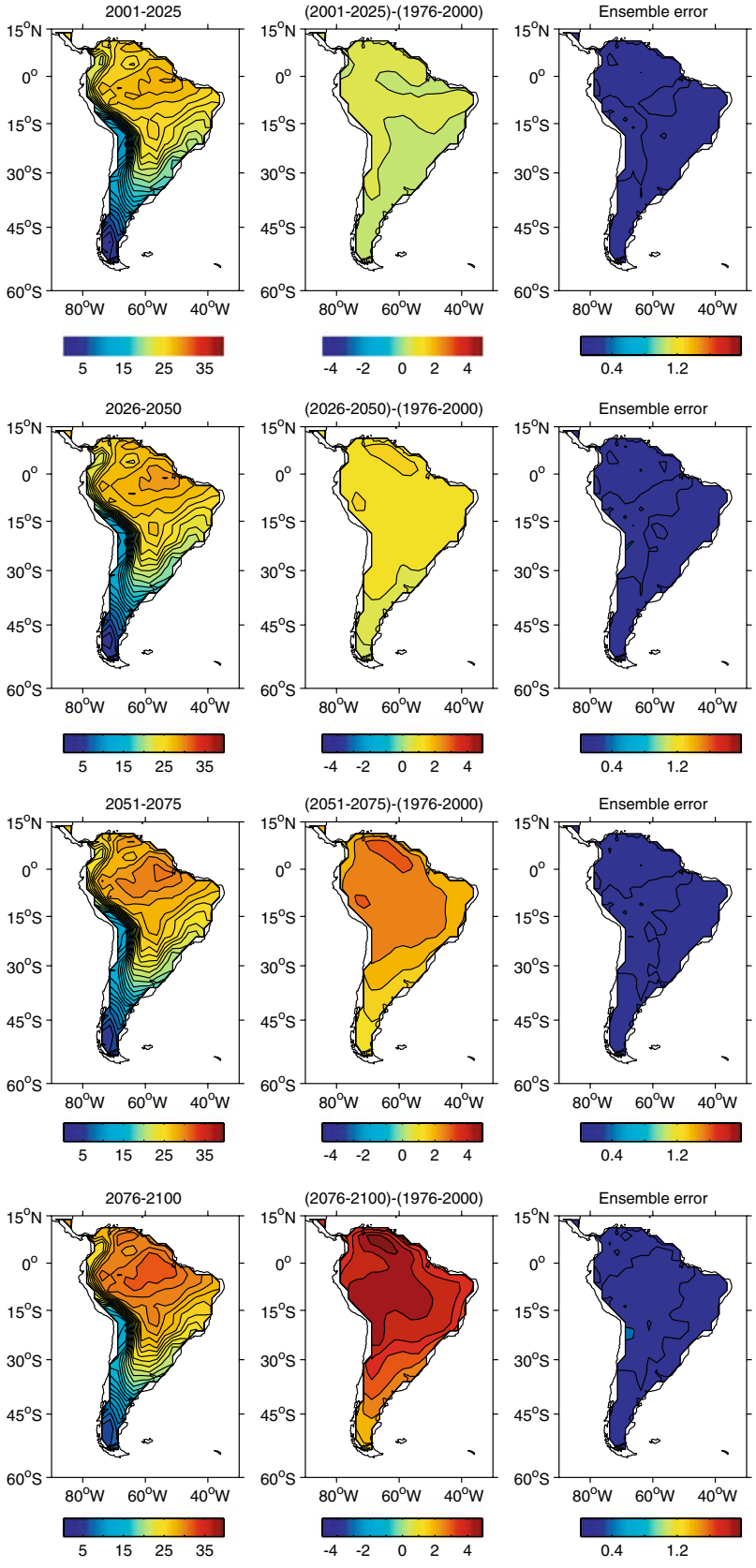


Fig. 13 From *top to down*, SRES A2 annual mean temperature projections for each period 2001–2025, 2026–2050, 2051–2075, and 2076–2100. *From left to right* Linear projection of the annual mean temperature (contours are every 1°C); differences between the linear projection and the 1976–2000 conditions (contours are every 0.5°C); Ensemble variance (contours are every 0.2°C)

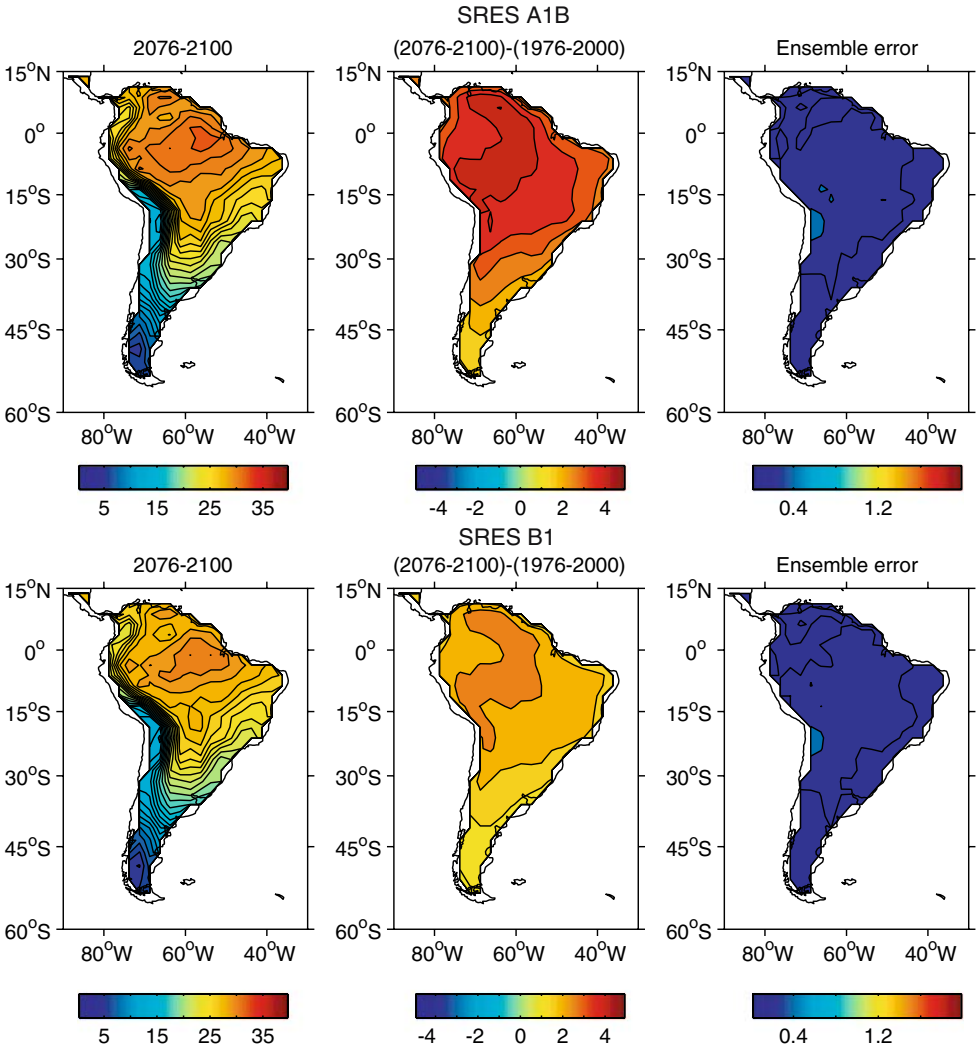


Fig. 14 From *top to down*, 2076–2100 projections respectively for SRES A1B and SRES B1. From *left to right*: same as Fig. 13

significant consequences for crops and health. In the last case, a milder winter may reinforce the risk of endemic Dengue in the southern part of the region, which may affect not only Brazil (as it already does) but also Uruguay and Argentina.

5.2.2 SRES A1B

Figure 17 displays late twenty-first century SRES A1B projection for each season. As can be observed, most of the features described earlier for SRES A2 would be valid for SRES A1B, and only differ in amplitude. Moreover, the regional indices show trends very similar to SRES A2 although of smaller amplitudes. On average, the warming trends computed as the difference between late twenty-first century and late twentieth century is about 80–90% of the value of those observed in SRES A2. It is worth pointing out that the major differences between SRES A2 and SRES A1B are observed in late the twenty-first century. Before that, the two scenarios are similar.

5.2.3 SRES B1

Figure 18 displays late twenty-first century SRES B1 projection for each season. As can be observed, most of the features described earlier for SRES A2 or SRES A1B are valid for SRES B1, and only differ in amplitude. Moreover, the regional indices show trends very similar to SRES A2, although with much smaller amplitudes. SRES B1 diverges from both SRES A2 and SRES A1B as soon as 2025. In the late twenty-first century, the trends suggested by SRES B1 are about half (50–60%) those observed in SRES A2. However, as pointed out earlier, the large-scale patterns are relatively similar to those observed in SRES A2 and SRES A1B.

6 Conclusion and discussion

A major challenge for the scientific climate community at the beginning of twenty-first century is to provide as accurate as possible an estimate of future climate conditions, according to potential economic scenarios of

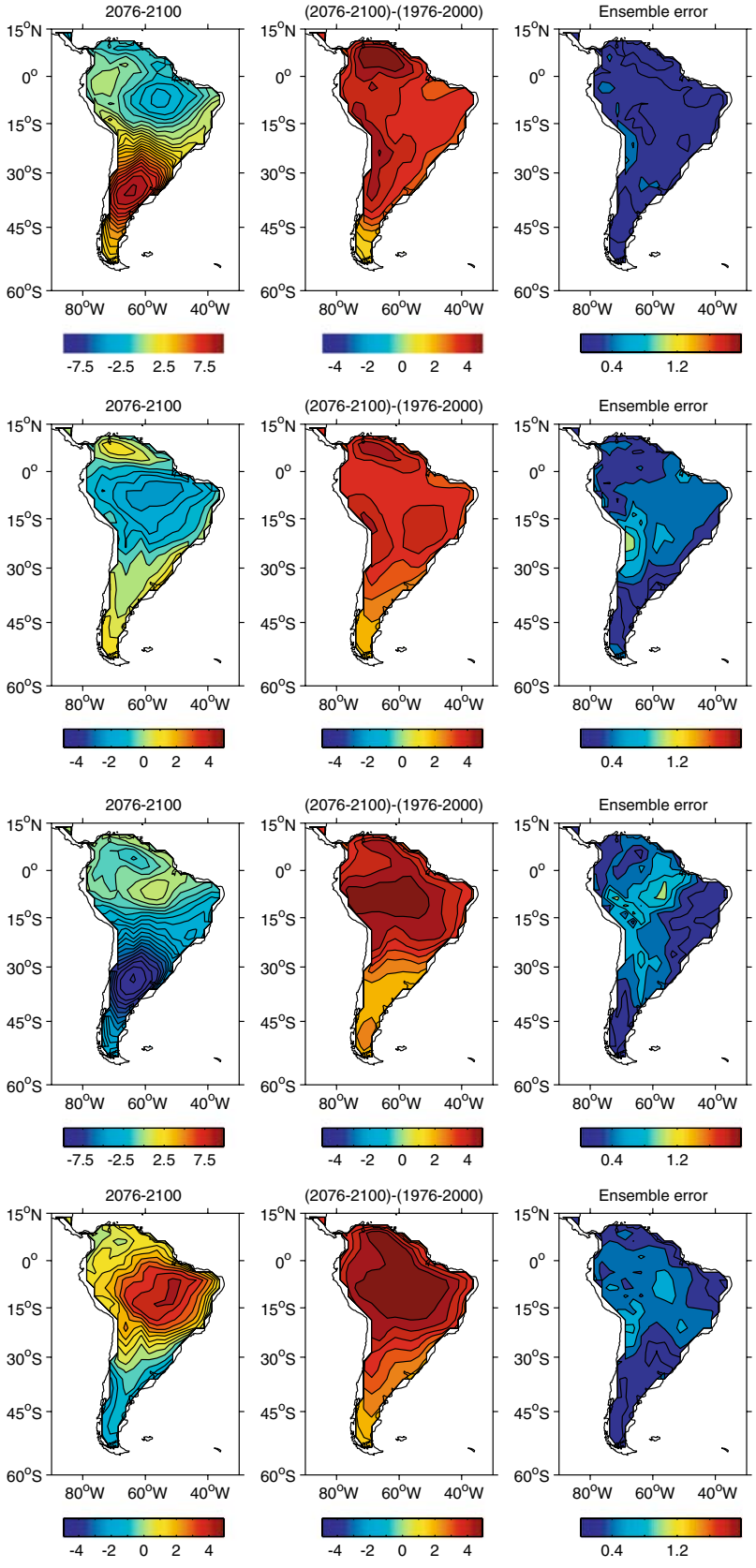


Fig. 15 From *top to down*, same as Fig. 13 but for each season (DJF, MAM, JJA, and SON)

evolution. The great international effort made by numerous climate centers to providing the entire community with ensembles of climate simulations for these scenarios is an important step toward that goal. There is no doubt that each climate model has skills in capturing certain aspects of the climate system mechanisms. It is

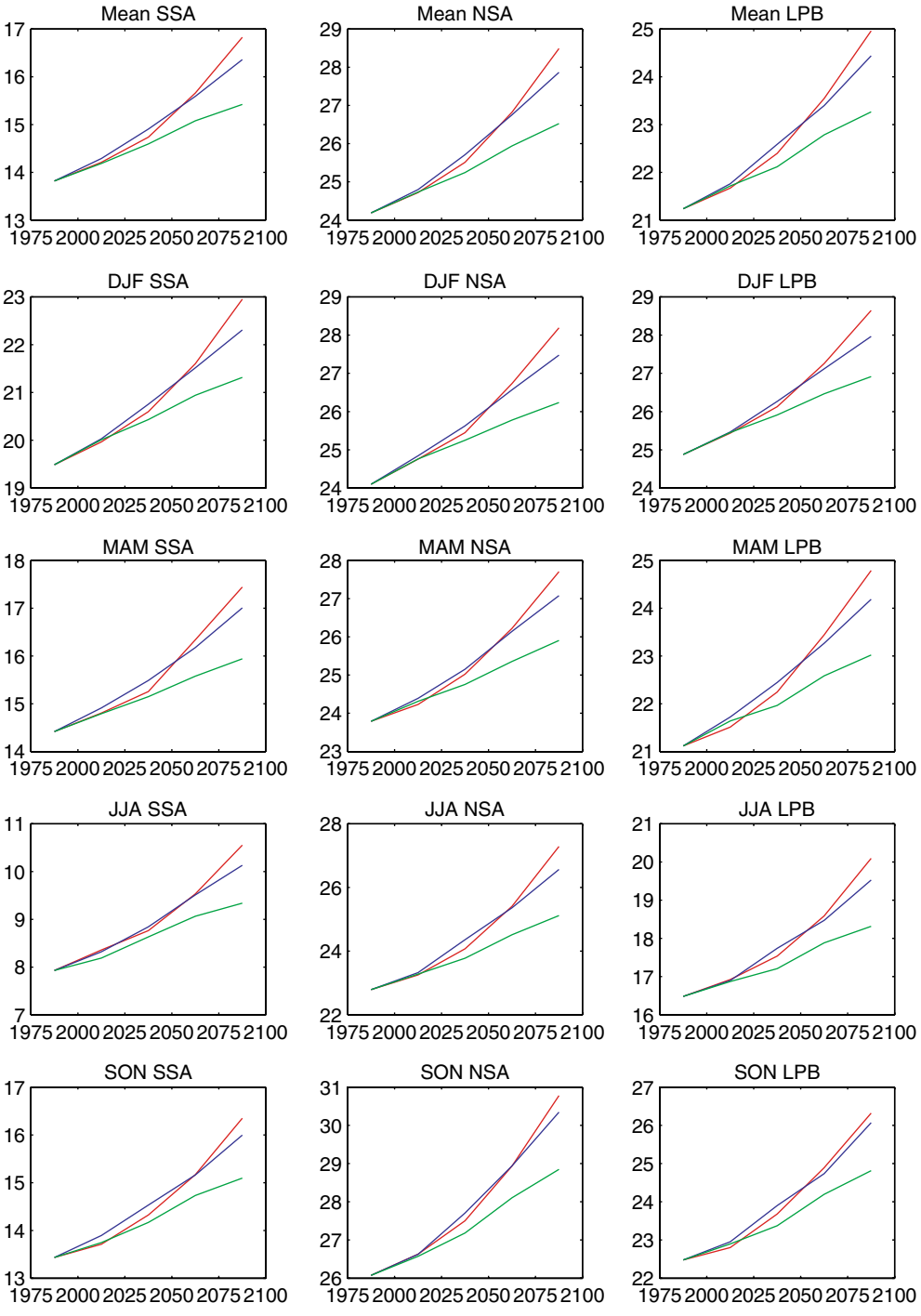


Fig. 16 Time evolution of annual and seasonal mean temperature in three different regions of South America. *NSA* stands for northern South America (north of 25°S). *SSA* stands for southern South America (south of 25°S). *LPB* stands for La Plata basin (simplified to a rectangular box extending from 15 to 35°S and from 65°W to the coast). SRES A2 projections are in *red*. SRES A1B projections are in *blue*. SRES B1 projections are in *green*

unlikely, however, that one specific model will capture all of them better than all other models. Therefore, multi-model/multi-ensemble analysis is a way for the future to take advantage of what each model represents best in order to optimize the projections of future climate change conditions. The present paper suggests a possible strategy to reach that goal. We focused on a small set of models (7) in order to describe the methodology, and we hope to be able to extend our work to a larger set of models in the future.

Our methodology is based on the use of neural networks, whose weights and hyperparameters are optimized through Bayesian statistics (see Appendix). As compared to other methods using Bayesian methods to combine model projections (e.g., Giorgi and Mearns 2002; Tebaldi et al. 2005), the use of neural networks provides a non-linear way to take into account model spatial biases in their simulation of present climate conditions. The major difficulty in using such a method is the optimization of the architecture of the neural

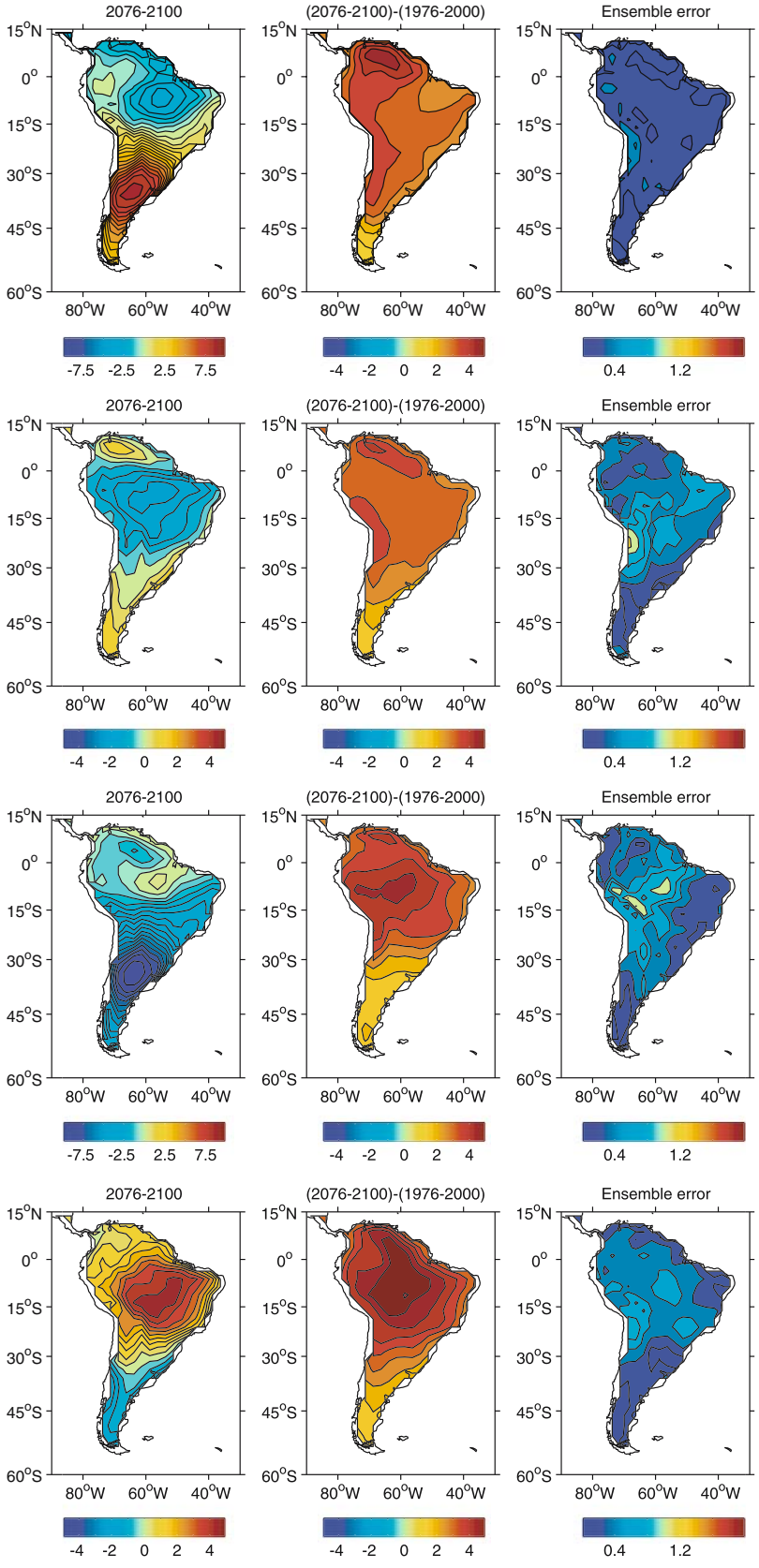


Fig. 17 Same as Fig. 15, but for SRES A1B projections

network, avoiding convergence in local minima, which would bias the optimal IPCC combination. Moreover, it is important to take into account different architectures, which may not be significantly different in representing present-day conditions, but may differ when projecting twenty-first century conditions. We took great care in

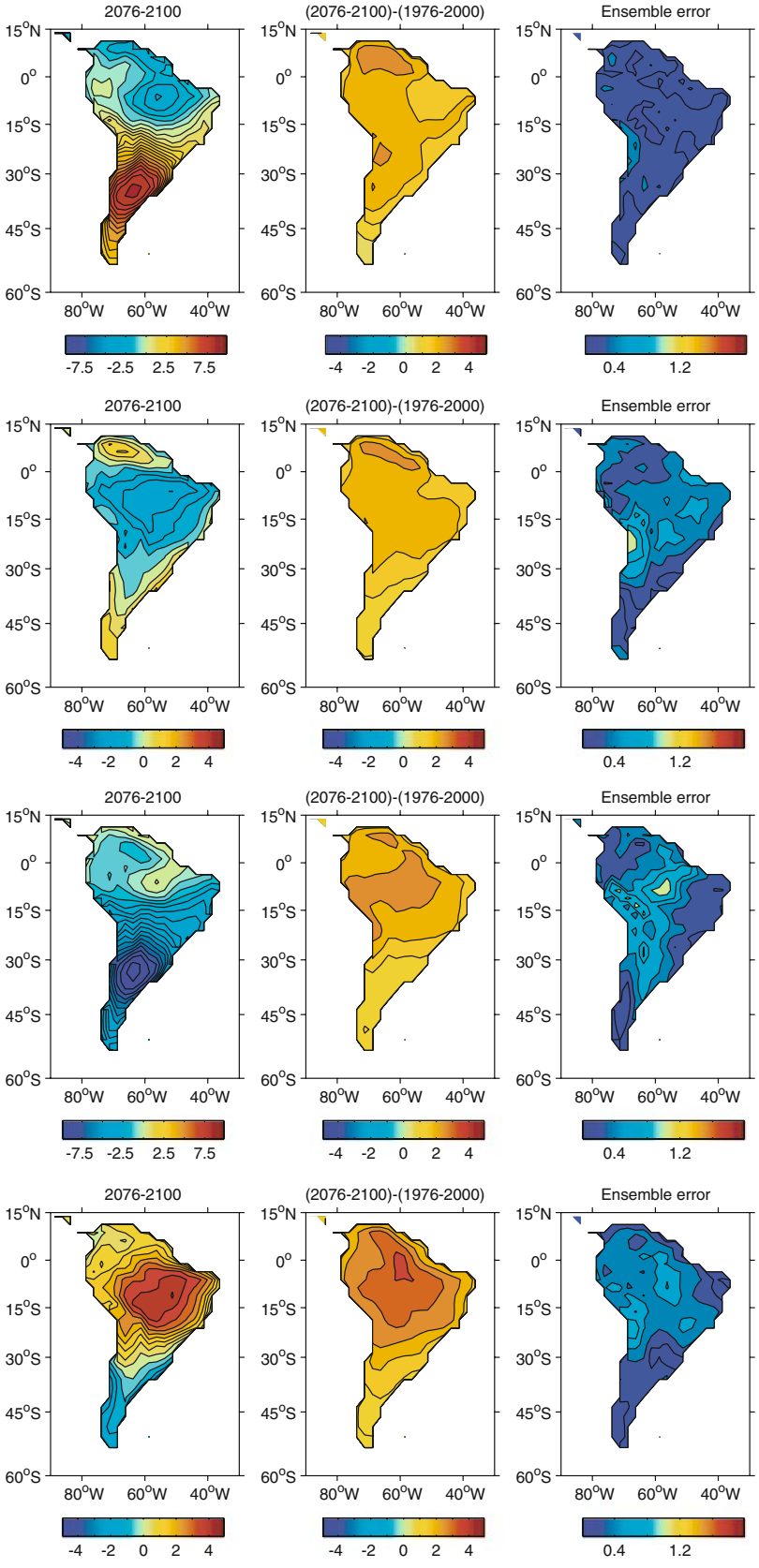


Fig. 18 Same as Fig. 15, but for SRES B1 projections

ensuring that the method fits present-day conditions for correct reasons (i.e., that a mixture of models is actually able to compensate for their bias, and avoids an over-fitting).

One of the outputs of the method is hyperparameter computation, called in the paper, MWI, which can be interpreted, by analogy to linear fitting methods, as normalized weights of a linear mixture of models. The MWI is not a quality criterion for models; it indicates, which models will contribute more to the transfer function represented by the neural network when mixing the models. Since we analyzed the entire South American continent, a model may have a weak MWI, but may have very good skill in simulating temperature in a sub region of South America. In order to use the MWI as a quality criterion one would have to divide the globe into climate-coherent regions, before performing the analysis. This was beyond the scope of the present study.

A major difficulty in using neural networks for climate change determining the network's skill to extrapolate. A comparison between MLP projection and a linear ensemble projection based on the MWIs allowed us to determine that the MLP penalizes future climate change projections (i.e., the MLP projections are of smaller amplitudes than any model or their linear combination). The ratio between MLP and linear projections (also called confidence level) is sensitive to two factors: the bias and the divergence criteria.They represent respectively the error between the linear combination and present-day climate conditions and the variance between the models. The largest confidence levels are observed on the Atlantic coast from the tropics to 35°S and the Pacific coast from Colombia to 15°S. In the other regions, the confidence level is low and can drop to zero in Chile between 15 and 30°S.

In conclusion, when applied to temperature, the neural network approach, using a Bayesian statistics for optimization, makes it possible to compute the optimal set of weights for a linear combination of the IPCC models, and a penalizing function or probability that such a change occurred, based on the present-climate model biases and their projection dispersion. Therefore, we focused on the linear ensemble projection, although the reliability of the results depends on the confidence level displayed in Fig. 10.

When projecting future climate conditions, we found that the three scenarios (A2, A1B, and B1) show similar patterns and differ only in amplitude, confirming results obtained by Ruosteenoja et al. (2003). However, SRES A1B differ from SRES A2 mainly in the late twenty-first century reaching about 80–90% amplitude (respective) compared to SRES A2. SRES B1, however, diverges from the other two scenarios as soon as in 2025. In the late twenty-first century, SRES B1 displays about half the amplitude of SRES A2.

Spatially, our major findings in SRES A2 for the end of the twenty-first century are that tropical South America may warm up by about 4°C with larger amplitudes over the Chilean and Peruvian coasts, the central Amazon and the Colombia–Venezuela–Guyana region. In the southern part of the continent, the warming could reach about 2–3°C. However, as pointed out before, the method indicates a large uncertainty (the confidence level is close to zero in the southern tip of South America, see Fig. 10). Interestingly, this annual mean temperature trend is modulated by the seasonal cycle in contrasted ways in each sub region. In SSA, the amplitude of the seasonal cycle would increase, while in NSA the amplitude of the seasonal cycle would be reduced. The reduction of the winter–summer contrasts together with a significant warming trend may induce strong impacts in these regions. In particular, diseases such as Dengue, which are vector-borne (Degallier et al. 2005), depend strongly on how long the mosquitoes live and how they survive cold winter conditions. In a much warmer climate than the one projected, it is likely that changes in winter conditions may increase the risk of Dengue developing to the south of its actual position. The study of such impacts in South America is under analysis in the framework of the European CLARIS Project.

Acknowledgements We wish to thank the Institut de Recherche pour le Développement (IRD), the Institut Pierre-Simon Laplace (IPSL), the Centre National de la Recherche Scientifique (CNRS; Programme ATIP-2002) for their financial support crucial in the development of the authors' collaboration. We are also grateful to the European Commission for funding the CLARIS Project (Project 001454) within whose framework the present study was undertaken. We are grateful to the University of Buenos Aires and the "Department of Atmosphere and Ocean Sciences for welcoming Jean-Philippe Boulanger. We thank Tim Mitchell and David Viner for providing the CRU TS2.0 datasets. Finally, we thank the European project CLARIS (http://www.claris-eu.org) for facilitating the access to the IPCC simulation outputs. We thank the international modeling groups for providing their data for analysis, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) for collecting and archiving the model data, the JSC/CLIVAR Working Group on Coupled Modelling (WGCM) and their Coupled Model Intercomparison Project (CMIP) and Climate Simulation Panel for organizing the model data analysis activity, and the IPCC WG1 TSU for technical support. The IPCC Data Archive at Lawrence Livermore National Laboratory is supported by the Office of Science, U.S. Department of Energy. Special thanks are addressed to Alfredo Rolla for his strong support in downloading all the IPCC model outputs.

7 Appendix: The multi-layer perceptron (MLP)

7.1 General description

The MLP is probably the most widely used architecture for practical applications of neural networks (Nabney 2002). From a computational point of view, the MLP can be described by a set of functions applied between different elements (neurons) using relatively simple arithmetic formulae, and a set of methods to optimize these functions based on a set of data. In the present study, we will only focus on a two-layer network architecture (Fig. 19). Its simplest element is called a neuron and is connected to all the neurons in the upper

layer (either the hidden layer if the neuron belongs to the input layer or the output layer if the neuron belongs to the hidden layer). Each neuron has a value, and each connection is associated to a weight (Fig. 20).

As shown in Fig. 19, in the MLP case we considered, the neurons are organized in layers: an input layer (the values of all the input neurons except the bias are specified by the user), a hidden layer and an output layer. Each neuron in one layer is connected to all the neurons in the next layer. More specifically, defining the input vector $(\xi_i)_{i=1,I}$, the first layer of the network forms H linear combinations (H is the number of neurons in the hidden layer) of the input vector to give the following set of intermediate activation variables:

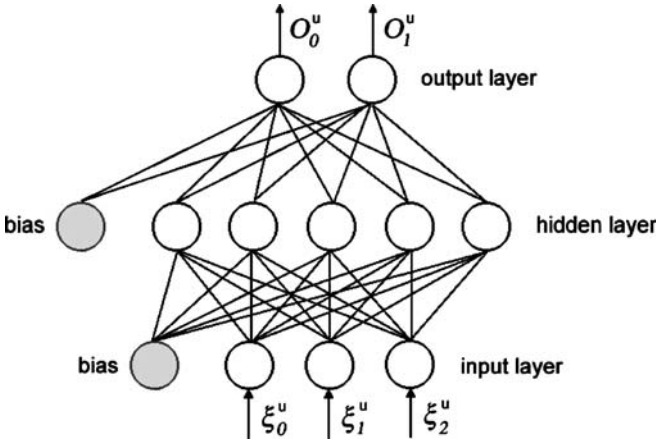


Fig. 19 Schematic representation of a two-layer MLP. ξ_u is the set of input value and o^u is its corresponding output (here we represent a specific case with a three-value input vector and a two-value output vector). The units or neurons called bias are units not connected to a lower layer. Their values are always equal to -1. They actually represent the threshold value of the next upper layer

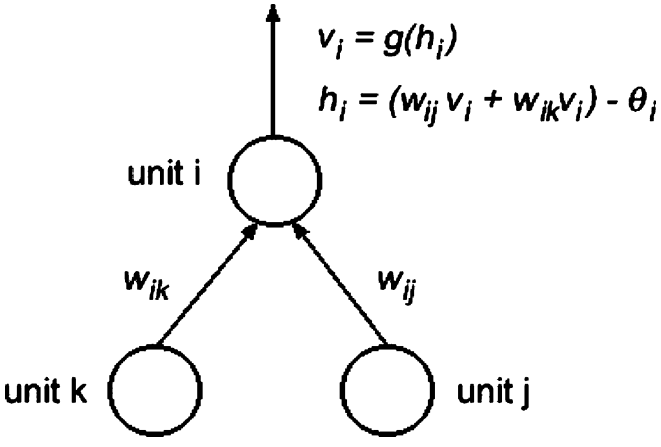


Fig. 20 In our case, each unit in a certain layer is connected to all the units in the lower layer. Each connection is associated to a specific weight, which value is optimized during the learning phase. θ is the bias value

$$h_j^{(1)} = \sum_{i=1}^{I} w_{ji}^{(1)} \xi_i + b_j^{(1)} \quad j = 1, \ldots, H,$$

where $b^{(1)}{}_j$ corresponds to the bias of the input layer. Then, each activation variable is transformed by a non-linear activation function, which in most cases (including ours), is the hyperbolic tangent function (tanh): $v_j = \tanh(h_j^{(1)}), j = 1,..., H$. Finally, the v_j are transformed to give a second set of activation variables associated to the neurons in the output layer:

$$h_k^{(2)} = \sum_{j=1}^{H} w_{kj}^{(2)} v_j + b_k^{(2)} \quad k = 1, \ldots, O,$$

where $b^{(2)}{}_k$ corresponds to the bias of the hidden layer. In most cases (including ours), the activation variables are associated to each neuron of the output layer through the linear function: $y_k = h^{(2)}{}_k$. Other more complex functions may be used according to the problem under consideration.

The weights and biases are initialized by random selection from a zero mean, isotropic Gaussian unit variance where the variance is scaled by the fan-in of the hidden or output units as appropriate. During the training phase, the neural network compares its outputs to the correct answers (a set of observations used as output vector), and it adjusts its weights in order to minimize an error function. In our case, the weights and biases are optimized by back-propagation using the scaled conjugate gradient method.

This architecture is capable of universal approximation and, given a sufficiently large number of data, the MLP can model any smooth function. Finally, the interested reader can find an exhaustive description of the MLP network, its architecture, initialization and training methods in Nabney (2002). Our study made use of the Netlab software (Nabney 2002).

7.2 Bayesian approach for selecting the "best" MLP architecture

When optimizing a model to the data, it is usual to consider the model as a function such as: $y = f(x, w) + \varepsilon$, where y are the observations, x the inputs, f the model, w the parameters to optimize (or the weights in our case) and ε the remaining error (model-data misfit). The more complex the model to fit (i.e., the number of parameters), the smaller the error, with the usual drawback of overfitting the data by fitting both the "true" data and its noise. Such an overfit is usually detected due to a very poor performance of the model on unseen data (data not included in the training phase). Therefore, optimizing the model parameters through minimizing the residual ε may actually lead to a poor model performance. One way to avoid such a problem is to consider also the errors in the model parameters. The use of a Bayesian approach is very helpful to address such an issue.

Although two kinds of Bayesian approaches have been demonstrated to be effective (Laplace approximation and Monte Carlo techniques), we will only consider the first one. Nabney (2002) offers an exhaustive discussion of this subject. For the reader to understand our approach, we believe the following summary is important.

First of all, following the same notations as in Nabney (2002), let's consider two models M1 and M2 (in our case two MLPs which only differ in the number of neurons in the hidden layer and M2 having more neurons than M1). Using Bayes' theorem, the posterior probability or likelihood for each model is:

$$p(M_i|D) = \frac{p(D|M_i)p(M_i)}{p(D)}.$$

Without any a priori reason to prefer any of the two models, the models should actually be compared considering the probability $p(D|M_i)$, which can be written (MacKay 1992) as $p(D|M_i) = \int p(D|w,M_i)p(w|M_i))\mathrm{d}w$. Considering that for either model, there exists a best choice of parameters for which the probability is strongly peaked, then the previous equation can actually be simplified:

$$p(D|M_i) \approx p(D|\hat{w}_i, M_i)p(\hat{w}_i|M_i)\Delta\hat{w}_i^{\text{posterior}},$$

where the last term represents the volume (in the space of the parameters) when the probability is uniform. Assuming that the prior probability $p(\hat{w}_i|M_i)$ has been initialized so that it is uniform over a certain volume of the prior parameters, we can rewrite the previous equation as:

$$p(D|M_i) \approx p(D|\hat{w}_i, M_i)\left(\frac{\Delta\hat{w}_i^{\text{posterior}}}{\Delta\hat{w}_i^{\text{prior}}}\right).$$

The new equation is the product of two terms evolving in opposite directions as the complexity of the model increases. The first term on the right-hand side increases (i.e., the model-data misfit decreases) as the model complexity increases. The second term is always lower than 1 and is approximately exponential with parameters (Nabney 2002), which penalizes the most complex models. In conclusion, if this is taken into account, the weight uncertainty should reduce the overfitting problem. We will now explain how this can be done.

For a given number of units in the hidden layer, an optimum set of weights and biases can be calculated using the maximum likelihood to fit a model to data. In such a case, this optimum set of parameters (weights and biases) is the one, which is most likely to have generated the observations. A Bayesian approach (or quasi-Bayesian approach due to difficulties in using Bayesian inference caused by the non-linear nature of the neural networks) may be valuable to infer these two classes of errors: model-data misfit and parameter uncertainty.

According to Bayes' theorem, for a given MLP architecture, the density of the parameters (noted w) for a given dataset (D) is given by:

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}.$$

In a first step, let's only consider the terms depending on the weights. The negative log likelihood is given by $E = -\,\mathrm{Log}(p(w|D)) = -\,\mathrm{Log}\,(p(D|w)) - \mathrm{Log}\,(p(w))$.

The likelihood p(D|w) represents the model-data fit error, which can be modeled by a Gaussian function of the form:

$$p(D|w) = \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left(-\frac{\beta}{2}\sum_{n=1}^{N}\{f(x_n, w) - y_n\}^2\right) = \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left(-\frac{\beta}{2}E_D\right),$$

where β represents the inverse variance of the model-data fit error, f is the MLP function of the inputs (x_n) and weights (w), and y_n are the observations.

The requirement for small weights (i.e., avoiding the overfitting) suggests a Gaussian distribution for the weights of the form:

$$p(w) = \left(\frac{\alpha}{2\pi}\right)^{W/2} \exp\left(-\frac{\alpha}{2}\sum_{i=1}^{W} w_i^2\right) = \left(\frac{\alpha}{2\pi}\right)^{W/2} \exp\left(-\frac{\alpha}{2}E_W\right),$$

where α represents the inverse variance of the weight distribution. α and β are known as hyperparameters. Therefore, in order to compare different MLP architectures, we need first to optimize the MLP weights, biases, and hyperparameters for any given architecture. Such an optimization can be made using the evidence procedure, which is an iterative algorithm. Here again, we refer the reader to Nabney (2002). Briefly, if we consider a model to be determined (for any given architecture) by its two hyperparameters, we can write (as previously) that two models may be compared through their respectively maximized evidence $p(D|\alpha, \beta)$, which log evidence can be written in the form:

$$\ln p(D|\alpha,\beta) = -\alpha E_W - \beta E_D - \frac{1}{2}\ln|A| + \frac{W}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{W+N}{2}\ln(2\pi),$$

where A is the Hessian matrix of the total error function (function of α and β).

Based on the previous equation, the evidence procedure is used to optimize the weights and hyperparameters for any given architecture, and the model optimized log evidence is calculated.

7.3 Model weight indices

Interestingly, the concept of hyperparameters introduced previously can actually be generalized by assigning a separate hyperparameter to each input neuron.

The hyperparameter represents the inverse variance of the weights fanning out from the input neuron to the hidden neurons. A high hyperparameter value means that the weights are small, close to zero and therefore that the corresponding input is less important. Therefore, the hyperparameter is indicative of the importance of the input to the rained output. Based on that information, we introduced a MWI, which is defined as the variance of the weights fanning out from a neuron (i.e., the inverse of the model input hyperparameter) normalized by the sum of the weight variances of all the model inputs. The MWI is comprised between 0 and 1, and indicates the relative importance of the different models to the trained output. By analogy, each MWI can be compared to a linear weight applied to each model when combining them linearly. We will show that, although the MLP represents better observations than any linear combination of the models, a linear combination based on the MWIs has also a certain skill in reproducing the observations. The linear combination of models based on MWIs is a simplified linear version of the neural network.

References

Allen MR, Stott PA, Mitchell JFB, Schnur R, Delworth TL (2000) Quantifying the uncertainty in forecasts of anthropogenic climate change. Nature 407:617–620

Boulanger J-P, Leloup J, Penalba O, Rusticucci M, Lafon F, Vargas W (2005) Low-frequency modes of observed precipitation variability over the La Plata basin. Clim Dyn 24:393–413. DOI 10.1007/s00382-004-0514-x

Coelho CAS, Pezzulli S, Balmaseda M, Oblas-Reyes FJD, Stephenson DB (2004) Forecast calibration and combination: a simple Bayesian approach for ENSO. J Clim 17:1504–1516

Collins WD et al (2005) The community climate system model, version 3. J Clim (in press)

Degallier N, Favier C, Boulanger J-P, Menkes C, Oliveira C, Rubens Costa Lima J, Mondet B (2005) Early determination of the reproductive number for vector-borne diseases: the case of dengue in Brazil. (in press)

Delworth et al (2005) GFDL's CM2 global coupled climate models—Part 1: Formulation and simulation characteristics. J Clim (in press)

Forest CE, Stone PH, Sokolov AP, Allen MR, Webster MD (2002) Quantifying uncertainties in climate system properties with the use of recent climate observations. Science 295:113–117

Giorgi F, Mearns LO (2002) Calculation of average, uncertainty range and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging" (REA) method. J Clim 15(10):1141–1158

Giorgi F et al (2001) Regional climate information: evaluation and projections. In: Houghton JT et al (eds) Climate change 2001: the scientific basis. Contribution of working group I to the 3rd assessment report of the intergovenmental panel on climate change, Chap 10. Cambridge University Press, Cambridge, pp 583–638

Gnanadesikan et al (2005) GFDL's CM2 global coupled climate models—Part 2: The baseline ocean simulation (in press)

Gordon C, Cooper C, Senior CA, Banks HT, Gregory JM, Johns TC, Mitchell JFB, Wood RA (2000) The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. Clim Dyn 16:147–168

Haak H et al (2003) Formation and propagation of great salinity anomalies. Geophys Res Lett 30: 1473. DOI 10.1029/2003GL17065

Johns TC, Carnell RE, Crossley JF, Gregory JM, Mitchell JFB, Senior CA, Tett SFB, Wood RA (1997) The second hadley centre coupled ocean–atmosphere GCM: model description, spinup and validation. Clim Dyn 13:103–134

Jones PD, Moberg A (2003) Hemispheric and large-scale surface air temperature variations: an extensive revision and an update to 2001. J Clim 16:206–223

MacKay DJC (1992) Bayesian interpolation. Neural Comput 4:415–447

Marsland et al (2003) The Max-Planck-Institute global ocean/sea ice modelwith orthogonal curvelinear coordinates. Ocean Model 5:91–127

Nabney IT (2002) Netlab. Algorithms for pattern recognition. Advances in Pattern Recognition. Springer, Berlin Heidelberg New York, pp 420

Nakicenovic N, Alcamo J, Davis G, de Vries B, Fenhann J, Gaffin S, Gregory K, Grbler A, Jung TY, Kram T, La Rovere EL, Michaelis L, Mori S, Morita T, Pepper W, Pitcher H, Price L, Raihi K, Roehrl A, Rogner H-H, Sankovski A, Schlesinger M, Shukla P, Smith S, Swart R, van Rooijen S, Victor N, Dadi Z (2000) IPCC special report on emissions scenarios. Cambridge University Press, Cambridge, pp 599

New MG, Hulme M, Jones PD (2000) Representing twentieth-century space–time climate variability. Part II: Development of 1901–1996 monthly grids of terrestrial surface climate. J Clim 13:2217–2238

Reilly J, Stone PH, Forest CE, Webster MD, Jacoby HD, Prinn RG (2001) Uncertainty in climate change assessments. Science 293(5529):430–433

Roeckner et al (2003) The atmospheric general circulation model ECHAM5 Report No. 349OM

Ruosteenoja K, Carter TR, Jylhä K, Tuomenvirta H (2003) Future climate in world regions: an intercomparison of model-based projections for the new IPCC emissions scenarios. The Finnish Environment 644. Finnish Environment Institute, 83 pp

Salas-Mélia D, Chauvin F, Déqué M, Douville H, Gueremy JF, Marquet P, Planton S, Royer JF, Tyteca S (2004) XXth century warming simulated by ARPEGE-Climat-OPA coupled system

Stouffer et al. (2005) GFDL's CM2 global coupled climate models—Part 4: Idealized climate response (in press)

Tebaldi C, Smith RL, Nychka D, Mearns LO (2005) Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multi-model ensembles (in press)

Wigley TML, Raper SCB (2001) Interpretation of high projections for global-mean warming. Science 293:451–454

Wittenberg et al (2005) GFDL's CM2 global coupled climate models—Part 3: Tropical Pacific climate and ENSO (in press)