Accurate and Fast Neural Network Emulations of Model Radiation for the NCEP Coupled Climate Forecast System: Climate Simulations and Seasonal Predictions

V. M. Krasnopolsky,

SAIC at NCEP/NOAA, and

Earth System Science Interdisciplinary Center (ESSIC), University of Maryland

M. S. Fox-Rabinovitz,

ESSIC, University of Maryland

Y. T. Hou, S. J. Lord,

NCEP/NOAA,

and A. A. Belochitski

ESSIC, University of Maryland

Corresponding author address:
Vladimir Krasnopolsky,
5200 Auth Rd.,
Camp Springs, MD 20746-4304;
Phone: 301-763-8000 x 7262, Fax: 301-763-8545,
E-mail: Vladimir.Krasnopolsky@noaa.gov

Submitted to Monthly Weather Review

July 2009

Revised version submitted

November, 2009

Abstract

The approach to accurate and fast calculating model physics using neural network emulations was previously developed by the authors for both long-wave and short-wave radiation parameterizations, or for the *full model radiation*, the most time-consuming component of model physics. It was successfully tested for a moderate resolution *uncoupled* NCAR CAM (Community Atmospheric Model) driven by climatological SST for a decadal climate simulation mode (Krasnopolsky et al. 2008a). In this study, the approach has been further developed and implemented into the NCEP *coupled* CFS (Climate Forecast System) with significantly *higher resolution* and *time dependent* CO_2. The higher complexity of NCEP CFS required introducing further adjustments to the neural network emulation methodology. Validation of the approach for the NCEP CFS has been done through a decadal climate simulation and seasonal predictions. The developed highly accurate neural network emulations of long-wave and short-wave radiation parameterizations are on average 16 and 60 times faster than the original/control long-wave and short-wave radiation parameterizations, respectively. A detailed comparison of parallel decadal climate simulations and seasonal predictions performed with the original NCEP model radiation parameterizations and with their neural network emulations is presented. Almost identical or close results for model prognostic and diagnostic fields have been obtained for the parallel decadal simulations and seasonal prediction that justifies the practical use of efficient neural network emulations of full model radiation for climate simulations and seasonal predictions.

1. Introduction

Calculation of model physics in a GCM (General Circulation Model) usually takes a very significant part of the total model computations. Evidently, this percentage is model dependent but full model radiation is the most time-consuming component of GCMs (e.g., Morcrette et al. 2007, 2008, Manners et al. 2009). In both climate modeling and NWP, the calculation of radiative transfer is necessarily a trade-off between accuracy and computational efficiency. There exist very accurate methods such as line-by-line procedures that could be employed ideally to calculate radiative fluxes for every grid-point at every time-step. If the radiation transfer were to be computed for every grid point and at all time steps, it would generally require as much CPU time or more than the rest of the model components, i.e., model dynamics and other physical parameterizations (Morcrette et al. 2008). Therefore a number of simplifications are usually made to reduce this cost to manageable levels.

For example, in the majority of modern radiative schemes, the correlated-k method (Lacis and Oinas 1991) is typically used to reduce the integration over wavelength by effectively binning wavelengths with similar absorption coefficients (k-terms). This simplification reduces greatly the number of monochromatic radiative transfer calculations required. The number of k-terms can be adjusted, which provides a trade-off between the accuracy and efficiency required for a given application. However, the correlated-k methods cannot be made sufficiently computationally efficient to allow calculations for every grid-point at every time-step.

To reduce the cost further, calculations are usually made at lower temporal and/or spatial resolutions. Quite drastic reductions in temporal resolution are often made (e.g., radiation calculations are made every one or three hours for the climate and global forecast models at NCEP and UKMO (Manners et al. 2008)). Between radiative transfer calculations major changes may occur in the radiative profiles (caused primarily by two factors: changes in clouds and changes in the angle of incident solar radiation) that are not represented. A reduced horizontal resolution approach (the radiative calculations are performed on a coarser grid with a following interpolation of the results to an original finer grid) is used to speed up radiation calculations at ECMWF (Morcrette et al. 2007, 2008). A reduced vertical resolution approach (the full radiation is calculated at every other vertical level and interpolated on the intermediate levels) is used in the Canadian operational Global Environmental Multiscale model (e. g. Coˆte´ et al. 1998a, 1998b). Such approaches reduce horizontal or vertical variability of radiation fields. Thus, these approaches may reduce the accuracy of a model's radiation calculation and its spatial or/and temporal consistency with other parts of model physics and with model dynamics, which may, in turn, affect negatively the accuracy of climate simulations and weather predictions.

Such a situation is an important motivation for developing new alternative numerical algorithms that provide faster calculations of model physics while carefully preserving their accuracy. Two techniques have been proposed to improve temporal and spatial resolution of radiation calculations: (1) the technique that improves interpolation of the radiative calculations from the coarse grid to the fine one (Morcrette et al. 2008) or

improve radiative calculations between the time steps for which full radiative calculations are performed (Venema 2007, Manners et al. 2008), and (2) the technique that introduces either new fast radiation *parameterizations* (Chevallier et al. 1998, 2000) or accurate and fast *emulations* of existing radiation schemes and parameterizations (Krasnopolsky 1997, Krasnopolsky et al. 2005a, 2008a) that can be used in a model at each grid point and at each time step instead of original slow radiative calculations.

A fast neural network (NN) based long wave radiation parameterization NeuroFlux (Chevallier et al. 1998, 2000) has been developed and tested in the ECMWF model. The NeuroFlux approach has a limited application (as discussed in Krasnopolsky et al. 2005b) because it has been developed for a particular formulation (Washington and Williamson 1977) of the long wave radiation physics only. Also, because of NeuroFlux's suboptimal design (as discussed in Krasnopolsky et al. 2005b), at vertical resolution of 60 layers and more, both accuracy and rapidity of NeuroFlux cannot be achieved simultaneously (Morcrette et al. 2008). Consequently, the NeuroFlux is used only for the 4D-Var linearized physics (Janiskova et al., 2002) for which the accuracy requirements are less stringent.

In our previous studies (Krasnopolsky et al. 2005a, 2008a, Krasnopolsky and Fox-Rabinovitz 2006a, b) we demonstrated that the neural network emulation approach can be successfully used to speed up significantly (by one to two orders of magnitude) the calculations of model radiation while providing a sufficient accuracy of decadal (50-year) climate simulations. We also demonstrated that this approach is a generic one; namely it

can be used not only for emulating any formulation of the long wave radiation physics but also for emulating any formulation of short wave radiation physics.

In the previous studies (Krasnopolsky et al. 2005a, 2008a), we used a moderate resolution NCAR (National Center for Atmospheric Research) CAM (Community Atmospheric Model), coupled with a land model, with the T42 (~3 degree) horizontal resolution and 26 vertical levels (T42L26). In that study, CAM was driven by the climatological sea surface temperature (SST) forcing (with no ocean model coupled).

It this study we apply the NN emulation approach to the higher complexity NCEP CFS (Climate Forecasting System) (Saha et al. 2006), which required further development of the neural network emulation methodology. We demonstrate that the NN emulation approach for model radiation can be successfully applied to the *significantly higher resolution coupled ocean-atmosphere-land-ice* model with *time dependent* CO_2. The atmospheric part of CFS has spectral T126 horizontal resolution and 64 vertical levels (T126L64); it is coupled with the 40-level interactive MOM4 ocean model, with a state-of-the-art 3D land model, and with an ice model.

In Section 2, we briefly describe the coupled NCEP CFS. In Section 3, the extended NN emulation methodology and developed NN emulations for the NCEP CFS long-wave radiation (LWR) and short-wave radiation (SWR) are briefly described in terms of their design, accuracy, and computational performance. In Section 4, the results of the parallel decadal model simulations and seasonal predictions, one using both LWR and SWR NN

emulations for calculation of full model radiation (the NN run) and the other using the original model radiation (the control run) are compared in terms of similarity of their spatial and temporal variability characteristics. Section 5 contains conclusions.

2. The NCEP Climate Forecast System

The operational NCEP CFS is described in detail in Saha et al. (2006) and the references therein. The coupled NCEP CFS version (being tested for the operational use) used in our study incorporates: the NCEP GFS (Global Forecast System) 64-level atmospheric model, the 40-level interactive MOM4 ocean model, the interactive Noah land model with four soil levels with improved treatment of snow and frozen soil, an interactive sea ice model with fractional ice cover and depth allowed, a sub-grid scale mountain blocking, a new seasonal climatological aerosol treatment, a historical CO_2 database from global observations collected by the World Meteorological Organization, a variable solar constant database, and historical stratospheric volcanic aerosol distributions (Sato et al., 1993).

The NCEP GFS model is a mature, state-of-the-art spectral atmospheric GCM (AGCM) used in operational medium-range weather forecasts. The operational GFS version has a variable horizontal spectral resolution of up to T382 or ~ 38 km. The hybrid sigma-pressure coordinate and a conservative finite-difference scheme are used in the vertical domain. The operational model is run with 64-layer vertical resolution between the surface and 0.27 hPa (about 60 km). The current version is implemented in the Message Passing Interface parallel environment. The GFS incorporates parameterizations of a

variety of physical processes important in the troposphere and stratosphere and implemented in a "plug compatibility" format that facilitates model development. Its radiation components contain a GCM version (v2.3) of the Rapid Radiative Transfer Model (RRTM) for LWR (hereafter referred to as RRTMG-LW) developed at AER Inc. (e.g. Mlawer et al., 1997; Iacono et al., 2000), and a SWR based on Chou's parameterization scheme (Hou et al., 2002; Chou and Suarez, 1999). In the new coupled CFS used in this study the SWR of the operational GFS has been replaced by a GCM version (v2.3) of the AER's RRTM SWR (hereafter referred to as RRTMG-SW) (e.g. Clough et al., 2005) to improve the accuracy of SWR calculation.

3. NN emulations for the NCEP CFS radiation

3.1 Background Information on NCEP CFS LWR and SWR

Radiation (LWR and SWR) parameterizations in an atmospheric model calculate radiation fluxes and heating rates in the earth-atmospheric system. The RRTMG-LW in the new CFS model employs a computationally efficient correlated-k method for radiative transfer calculations. It contains 16 spectral bands with various number of quadrature points (g-points) in each of the bands that sums up to a total of 140 g-points (e.g., Mlawer et al., 1997, Iacono et al., 2000). Active gas absorbers include H_2O, O_3, CO_2, CH_4, N_2O, O_2, and four types of halocarbons (CFCs). A maximum-random cloud overlapping scheme is used for cloudy sky radiative transfer, and a climatological aerosol scheme provides the global distribution of aerosol optical depth. In this study, a one-hour frequency of radiation calculation is applied to both SWR and LWR. In the current version of the LWR parameterization the level of atmospheric CO_2 and its time

dependence is presented by its global mean value that increased from 350 to 380 ppmv during the period of model integration used in this study (1990 to 2006).

Beside the RRTMG-LW, which is a faster member of the RRTM LWR family, we have also experimented with another version of the RRTM LWR (hereafter as RRTMF-LW) in this study. The RRTMF-LW is based on AER's RRTM-LW v3.0. It uses a full 16 g-points in each of the 16 spectral bands that add to a total of 256 vs. the reduced total of 140 in the faster RRTMG-LW. Unlike the diffusivity approach (one zenith angle of about 53°) in the faster RRTMG-LW, the RRTMF-LW uses multi-angle radiance integration over a hemisphere to yield better accuracy (we set it at 3 angles in the study). As a result, the RRTMF-LW is about five times slower than the RRTMG-LW in exchange for improved accuracy (Mlawer et al., 1997).

The SWR parameterization used in the new CFS is a modified version of AER's RRTMG-SW (v2.3) (Clough et al., 2005). It contains 14 spectral bands with various numbers of g-points in each of the bands to a total of 112. RRTMG-SW uses a fast two-stream radiative transfer scheme, and supports sophisticated absorption and scattering processes by clouds, aerosols, and absorbing gases (H_2O, O_3, CO_2, CH_4, N_2O, O_2). Thus, in the current version of the SWR parameterization the level of atmospheric CO_2 and its time dependence is presented by the entire 3-D CO_2 field that changes with time in accordance with the change of the mean CO_2 level that increased from 350 to 380 ppmv during the period of model integration used in this study (1990 to 2006).

Although both RRTMG-LW and RRTMG-SW are built with fast computation schemes designed for GCM applications, they still represent the most time-consuming physics in the NCEP CFS model. The percentage of the total model computation time used by model physics and by radiation (LWR and SWR) vary depending largely on the model horizontal and vertical resolution, the time step, the frequency of radiative calculations, and the computing environment (e.g. the number of processors and threads). For example, in the new CFS configuration at the T126L64 resolution, with the new RRTMG-LW and RRTMG-SW both called every hour, the portion of the radiation computation time is about 57% of the total AGCM model computation time.

3.2 *Background Information on the NN emulation approach*

As we showed in our previous works (e.g., Krasnoposky et al. 2005a, Krasnopolsky 2007a) any parameterization of model physics can be emulated using NNs. NN is an analytical approximation that uses a family of functions like:

$$y_q = a_{q0} + \sum_{j=1}^{k} a_{qj} \cdot \phi(b_{j0} + \sum_{i=1}^{n} b_{ji} \cdot x_i); \quad q = 1, 2, \ldots, m \tag{1}$$

where x_i and y_q are components of the input and output vectors X and Y, respectively, a and b are fitting parameters, and $\phi(b_{j0} + \sum_{i=1}^{n} b_{ji} \cdot x_i)$ is a "neuron". The activation function ϕ is usually a hyperbolic tangent, n and m are the numbers of inputs and outputs respectively, and k is the number of neurons in (1). Definitions of NN terminology can be found in many places, for example in the book by Bishop (2006) and in the review paper by Krasnopolsky (2007a); however, eq. (1) is sufficient to understand the subject of

this paper. The numerical complexity of NN (1) can be well approximated by the number of NN weights, *a*'s and *b*'s in (1) (Krasnopolsky 2007a):

$$N_C = k \cdot (n + m + 1) + m \qquad (2)$$

The time, T_{NN}, required for estimating the NN (1) is directly proportional to the NN numerical complexity N_C,

$$T_{NN} = c \cdot N_C$$

with the coefficient of proportionality *c* depending mainly on a hardware and software environment of the computer used.

Obviously, the numerical complexity, N_C, increases linearly with the increase of model vertical resolution (the number of the vertical layers, *L*) because both *n* and *m* depend linearly on *L*. Thus as a result, the time required for estimating NN, T_{NN}, increases linearly with the increase of model vertical resolution. The time required for estimating the original parameterization, T_O, also increases with the increase of vertical resolution. For the original parameterization, the dependence of the calculation time on vertical resolution is strongly conditioned by the numerical scheme implemented. For example, calculation time increases for RRTMG-LW approximately linearly with the increase of *L* vs. an L^2 relationship for most of other types of LWR used in many models.

Thus, the dependence of the speedup, η, provided by a NN emulation on model vertical resolution is determined by the ratio of the two aforementioned calculation times:

$$\eta = \frac{T_O}{T_{NN}}$$

the time required for estimating the original parameterization, T_O and the time required for estimating the NN emulation, T_{NN}. Therefore, the change of the speedup η with the increase of model vertical resolution will strongly depend on the physical complexity of the original parameterization and on the numerical scheme implemented (see also the discussion at the end of Section 3.6).

The major goal for developing NN emulations for model physics is to obtain a sufficiently high accuracy for NN emulation with practically zero biases or systematic errors (calculated against the original model physics). This is a necessary condition for obtaining non-accumulating errors during long-term climate simulations which use developed NN emulations. *The choice of an optimal version of NN emulation is based on the accuracy, not on a speed-up of computation. All the NN emulations obtained provide a significant speed-up η anyway.* The most efficient and convenient way of developing NN emulations for model physics components is to develop a *single* NN for a model physics parameterization. Such an approach has been introduced, discussed, and applied in our research (e.g. Krasnopolsky et al. 2005a, b, 2008a).

3.3 NN Emulations for Full Model Radiation

The LWR and SWR parameterizations together comprise the full model radiation. The LWR and SWR parameterizations or the *full model radiation* for the NCEP CFS have been emulated using two NNs, one for LWR and another for SWR.

The input and output vectors for NNs, emulating the LWR or SWR parameterizations, include the same parameters as those of the input and output vectors for the original LWR or SWR parameterizations, respectively. For the LWR NN emulation, these input parameters are the following nine profiles: atmospheric pressure, temperature, specific humidity, ozone mixing ratio, total cloud fraction, cloud liquid water path, mean effective radius for liquid cloud, cloud ice water path, and mean effective radius for ice cloud. The time dependent CO_2 is presented, as in the original LWR parameterization, by its changing with time global mean. The LWR parameterization (and LWR NN emulation) output vectors consist of the profile of heating rates (HRs) and five radiation fluxes: the total sky outgoing LW radiation flux from the top layer of the model atmosphere (the outgoing LWR or OLR), the clear sky upward flux at the top of the model atmosphere, the total sky upward flux at the surface, the total sky downward flux at the surface, and the clear sky downward flux at the surface.

The NN emulation of the LWR parameterization includes all non-constant inputs of the original LWR (total 556 inputs; n = 556 in eq. (1)). It has the same outputs (total 69 outputs; m = 69 in eq. (1)) as the original LWR parameterization. We have developed several NNs, all of which have the same aforementioned inputs and outputs, with the number k changing from 50 to 200 in eq. (1). Varying k, the number of terms (or neurons) in eq. (1), allows us to demonstrate the dependence of the accuracy of approximation on this parameter as well as its convergence, and as a result, to provide a sufficient accuracy of approximation for the model (e. g. Krasnopolsky et al. 2005a).

The input vectors for the SWR parameterization include 55 vertical profiles: atmospheric pressure, temperature, specific humidity, ozone, CO_2, N_2O, O_2, and CH_4 volume mixing ratios, total cloud fraction, cloud liquid water path, mean effective radius for liquid cloud, cloud ice water path, mean effective radius for ice cloud, and three profiles (optical depth, single scattering albedo, and asymmetry parameter) for each of 14 different species of aerosols. The input vectors include also the solar zenith angle, the solar constant and the surface albedo for four different bands. The SWR parameterization output vectors consist of a vertical profile of heating rates (HRs) and nine radiation fluxes: three fluxes at the top layer of the model atmosphere (the total sky outgoing SW radiation flux, the total sky downward flux, the clear sky upward flux), four radiation fluxes at the surface (the total sky upward and downward fluxes and the clear sky upward and downward fluxes), and the downward (the total and clear sky) fluxes in the UV-B spectral band.

The NN emulations of the SWR parameterization have 562 inputs and 73 outputs. We have developed several NNs, with the number k changing from 50 to 200 in eq. (1). It is noteworthy that, as in the case of the NN emulation of LWR, the number of NN inputs is less than the number of input profiles multiplied by the number of vertical layers plus the number of relevant single level characteristics. Many input variables (e.g., almost all gases) have zero or constant values for the upper vertical layers, and for some gases the entire volume mixing ratio profile is a constant (obtained from climatological data).

To improve the accuracy of the approximation, these constant inputs were not used for NN training. Constant inputs (zero or nonzero) do not contribute to the functional input/output relationship and should not be used for development of NN emulations. Moreover, if they were used, they would introduce an additional noise (an approximation error). In addition, for SWR, 2688 inputs describing the optical depth, the single scattering albedo, and asymmetry parameters of 14 aerosol species were substituted by five inputs: $\cos(\tau)$, $\sin(\tau)$, $\cos(lon)$, $\sin(lon)$, and *lat*, where *lon* is the longitude, *lat* is the latitude, and $\tau = \frac{2 \cdot \pi}{T} \cdot q$, where q is the month of the year, and $T = 12$. Such a substitution is possible because in NCEP CFS aerosols are calculated using the specific humidity profiles and 3-D lookup tables composed of climatological monthly data, different for different months of the year. It means that the aerosol inputs are actually highly correlated, and, in terms of functional input/output dependences, the aerosol characteristics are the functions of *lat*, *lon*, τ, and the profile of specific humidity only. Since the profile of the specific humidity has been already included in NN SWR inputs, only the five aforementioned additional variables have to be included to allow NN to completely emulate the contribution of aerosols into SWR.

We would like to stress that not including the constant profiles as inputs into the NN emulation or reducing the number of highly correlated profiles (as in the case of aerosols) does not in any way diminish the accuracy of the NN emulation (it is demonstrated in Section 3.5 and 4.). All changing/non-constant inputs are included into the NN emulation. All constant inputs are included in the original parameterization when NCEP CFS has been used to generate the NN training data sets (see the next Section), and the

NN emulation receives information about them from the training data during the NN training process. From the practical point of view, if the values of constant inputs are changed or are made variable in the future, the NN emulation inputs will be adjusted correspondingly and NN will be retrained.

It was mentioned above that radiation parameterizations have two kinds of outputs: the heating rates and the radiative fluxes. They are not completely independent; there is a balance relationship between them (see Appendix 1). The outputs of original radiation parameterizations satisfy the balance relationship with high accuracy because the relationship is explicitly (or implicitly) included into the parameterizations. Obviously, the outputs of the NN emulations satisfy it only approximately but with high accuracy. A balancing procedure has been developed to balance NN outputs exactly. Table 1 shows that the balancing procedure does not practically affect the overall accuracy of LWR NN and only marginally improves the overall accuracy of SWR NN.

3.4 Generating data sets for NN training and validation

The NCEP CFS (T126L64) has been run for seventeen years to generate representative training data sets. The representative data set samples the atmospheric state variability adequately, i.e., it represents all possible states produced by the model as fully as possible (including the states introduced due to time dependent CO_2 concentration). All inputs and outputs of original LWR and SWR parameterizations have been saved for two days per month, i.e., for one day at the beginning and one day in the middle of the month, every three hours (eight times per day) to cover the annual and diurnal cycles. From each

of the three hourly global data set three hundred events (the set of input and output profiles) have been randomly selected. The obtained data set has been divided into three independent parts, each containing about 200,000 input/output vector combinations. The first part has been used for training, the second one for tests (control of overfitting, control of NN architecture, etc.), and the third part (an independent validation data set) has been used for validation of trained NN only. All approximation statistics presented in this section are calculated using this independent validation data set. The accuracy of the NN emulation, i.e., biases, rmse, etc., are calculated against the control (the original parameterization).

3.5 Bulk Approximation Error Statistics

To ensure a high quality of representation of the LWR and SWR processes, the accuracy of their NN emulations has been carefully investigated. The NN emulations have been validated against the original NCEP CFS LWR and SWR parameterizations. To calculate the error statistics presented in Tables 1 and 2 and in Fig.1, the original parameterizations and their NN emulations have been applied to the validation data set. Two sets of the corresponding HR profiles have been generated for both LWR and SWR. Total and level bias (or a mean error), total and level RMSE, profile RMSE or PRMSE, and σ_{PRMSE} have been calculated (see Appendix 2 and *Krasnopolsky 2007a* for the definition of these statistics).

< Table 1 >

Table 1 shows bulk validation statistics for the accuracy of approximation of heating rates (HR). It also shows the average computational performance or the speedup, which depends significantly on cloud situations as shown in Table 2, obtained for a single processer configuration for the best, in terms of both the accuracy and performance, developed NN emulations for the NCEP CFS LWR and SWR. For comparison, the information on the NN emulations for NCAR CAM LWR and SWR (Krasnopolsky et al. 2008a) is also presented in Table 1. Total statistics show the bias, RMSE (see eq. A2.1 in Appendix 2), PRMSE, and σ_{PRMSE} (A2.4) for the entire 3-D HR fields. Also, layer statistics (A2.2) for the top and bottom atmospheric layers are included to illustrate the accuracy of NN emulations in the areas of the increased non-linearity (Morcrette et al. 2008). Although the two models as well as their embedded radiation parameterizations are different, comparisons between NCAR CAM (with 26 vertical layers) and NCEP CFS (with 64 vertical layers) allow us to observe a general dependence of the NN accuracy on the model vertical resolution (see also error profiles shown in Fig.1).

< Figure 1 >

As can be concluded from Table 1 and Fig.1, NN emulations for both LWR and SWR handle very well the nonlinearity at the top of the atmosphere where biases and RMSEs are very small with RMSEs being even smaller than the total RMSE. At the bottom layer, the non-linearity does not cause significant increases in biases; the RMSEs increase about two times, but as compared with the total RMSE, remain sufficiently small.

It is noteworthy that the approximation errors are identified as being "sufficiently small" if they are of such a small magnitude that they have almost negligible impacts on model behavior as demonstrated in Section 4 for NCEP CFS and by Krasnopolsky et al. (2008a) for NCAR CAM. Only validation of NN emulations in parallel model runs allows us to make final conclusion about the sufficient smallness of the approximation errors.

It terms of the presented accuracy statistics, there are practically no differences between NCAR CAM with 26 vertical layers and NCEP CFS with 64 vertical layers. As shown in Fig.1, the entire vertical distributions of errors (for both LWR and SWR) are similar for these two models. Thus, the accuracy of our NN emulation approach does not depend significantly on vertical resolution of the model. It does depend on the vertical location of the atmospheric layer. The layer RMSE increases near the surface for both models.

Also, the NN complexity N_C (2) and average speedup η (how many times NN emulation is faster than the original parameterization) are shown in Table 1 (see the further discussion of the speedup η in Section 3.6 below). These characteristics complement our discussion on the dependence of the speedup on vertical resolution (see the end of Section 3.2). For the LWR parameterization, we see a significant decrease of the speedup for NCEP CFS with 64 vertical layers vs. NCAR CAM with 26 vertical layers although the LWR NN emulation for NCEP CFS is still 16 times faster than the original parameterization. For the SWR parameterization the opposite tendency is observed; that is, the speedup for NCEP CFS SWR NN is more than two times higher than that of NCAR CAM SWR NN.

These seemingly contradictory speedups for LWR and SWR emulations can be explained (as was mentioned above in Section 3.2) by the interplay of the two main contributing factors: the physical complexity of the radiation calculation itself (the number of treated species, spectral bands, parameterization schemes, etc.), and the dependence of the particular numerical scheme implemented in the radiative transfer on the number of vertical model layers. The results presented in Table 1 illustrate the fact that the numerical scheme implemented in the NCEP CFS RRTMG-LW parameterization is significantly more efficient (linear with respect to the number of vertical levels L) than that of the original NCAR CAM LWR parameterization (quadratic with respect to L). Thus, a smaller speedup factor is produced by the NN emulation for NCEP CFS LWR than that of NCAR CAM LWR.

The NCEP CFS's RRTMG-SW includes more spectral bands and g-points and uses more complex treatment for a larger variety of absorbing/scattering species; thus NN emulation shows a larger speedup value η than that of NCAR CAM SWR. In any case, our NN emulation approach is significantly less dependent (in terms of both the accuracy and speedup) on the increase of vertical resolution than the NN based LWR parameterization NeuroFlux for which at vertical resolution of 60 layers and more, both accuracy and speedup could not be achieved simultaneously (Morcrette et al. 2008). For our NN emulation approach, for the model with 64 vertical layers, the desired accuracy of the NN emulation could be achieved simultaneously with a significant speedup of ~ 16 times for the LWR and of ~ 60 times for the SWR parameterizations.

3.6 Estimation of Speedup

The radiatve transfer calculations take different time under different cloud conditions because of the different complexity of cloud-radiation interaction. We performed more accurate estimations of speedup separately for three different types of cloudiness: clear sky, three layer clouds, and a more complex cloud condition of deep convection. 3,000 profiles have been used for each test. The results for the calculation time and speedup are presented in Table 2. For a more complex cloud-radiation interaction (deep convection) the calculation of the original LWR and SWR parameterizations takes ~22% and ~57% more time respectively than for clear sky conditions. Obviously, the time of the NN radiation calculations does not depend on the cloud conditions. Thus, the speedup is significantly higher for the more complex cloud-radiation interaction.

< Table 2 >

As Table 2 shows, the average speedup presented in Table 1 is closer to the minimal speedup obtained under clear sky conditions. The results presented in Tables 1 and 2 were obtained in a sequential code-by-code comparison and represent adequately the situation when the model is run on a single processor. However, if we compare the control model run using the original parameterizations with the NN run when both runs use multiple processors and threads, the actual speedup will be significantly higher and closer to the maximum value shown in Table 2 because it will be determined by the slowest calculation, which is the deep convection condition. Radiation in the control run

for all other cloud conditions is calculated faster but the next integration time step will not start before the radiation calculations for the deep convection condition are completed; however, the time of the NN radiation calculations in the NN run does not depend on the cloud conditions. Thus, in the case of parallel calculations utilizing multiple processors and threads, in addition to a significant speedup, the use of the NN emulations in the model provides an additional advantage, namely it helps to achieve a significantly better load balance.

Using NN emulations *simultaneously* for LWR and SWR or for the full model radiation results in an overall significant, about 20 – 25% speedup of NCEP CFS climate simulations and seasonal predictions when both LWR and SWR are calculated every hour. The speedup η provided by NN emulations (see Table 1) can be also used for more frequent calculations of model radiation.

4. Validation of parallel decadal model simulations and seasonal predictions

In this section we present comparisons between two parallel 17-year NCEP CFS model runs: one using the original LWR and SWR parameterizations (the control run) and another one using their NN emulations. Both spatial and temporal characteristics of prognostic and diagnostic fields are compared for the parallel runs.

4.1 Measures for Assessment of the Impact of Using NN Emulations of Full Model Radiation

We show below the differences between the two parallel runs. To evaluate the NN induced changes, we compare them with such commonly used measures as observation errors or uncertainties of reanalysis. We show that the differences are smaller than these quantitative measures.

In order to emphasize how small the changes introduced by the use of NN emulations are, we also find it appropriate to use a measure derived from the model itself, namely, the model's internal variability. Because a GCM is an essentially nonlinear system, it may produce something like a "butterfly effect", that is a significant reaction/response even to small perturbations in the model or in the model computational environment (e.g. routine changes in computer hardware, operational system, compilers, libraries, etc.). Any, even infinitesimal change in model formulation, initial conditions or computational environment makes two model integrations diverge, with the effect that after the deterministic predictability is lost (which takes just weeks for the atmosphere, although longer for the ocean), the timing and location of weather patterns becomes essentially independent for the two integrations. Hence the two control model runs produced with the aforementioned small changes provide, in essence, two independent samples of the model's climatology, and their difference represents the model's internal variability.

Thus, we can state that the approximation error of NN emulation is negligible and, therefore, NN's accuracy is sufficient for the use in the model if the differences/changes introduced in the model results by using the NN emulation are of the same order of magnitude as the aforementioned model's internal variability.

To estimate the model's internal variability, we produced two control runs with the original NCEP CFS model configuration, i.e., without NNs. The first run was performed before and the second run after the routine changes (introduced quasi-regularly by system administrators) of the version of the FORTRAN compiler and libraries. Small differences between these two runs (which are similar to those due to changes in a computer operation system and/or in hardware (Moorthi 2009)) are shown below together with the differences between the parallel NN and control runs for comparison purposes, as an additional measure of the NN emulation accuracy. Presenting model's internal variability helps us to better evaluate the differences in climate simulations and seasonal predictions caused by using NN emulations for model radiation and to emphasize how small these differences are.

4.2 Comparison of Parallel Runs

4.2.1 Climate simulations

The results of 17-year (1990-2006) climate simulations performed with NN emulations for both LWR and SWR, i. e., for the full model radiation, have been validated against the parallel control NCEP CFS simulation using the original LWR and SWR. We analyze the differences between the parallel runs in terms of spatial (global) means as well as temporal characteristics.

< Figure 2 >

Let us discuss first the differences between the parallel simulations in terms of spatial and temporal radiation characteristics. The differences between the NN radiation and control runs and the differences between two control runs for zonal and time mean LWR and SWR fluxes are presented in Fig. 2. The upper row of Fig. 2 shows the differences for zonal and time mean top of atmosphere upward long (left panel) and short (right panel) wave fluxes (in W/m^2) for winter. The lower row of Fig. 2 shows the differences for zonal and time mean downward (left panel) and upward (right panel) surface long wave fluxes (in W/m^2). For the fluxes presented in Fig. 2, both the differences between the NN radiation and control runs and the differences between two control runs are small and similar by magnitude. They do not exceed 2-3 W/m^2, i.e., they are within observational errors and uncertainties of reanalysis (e.g. Kalnay et al. 1996, Kistler et al. 2001). The similarity of the differences by magnitude means that both the differences between the NN radiation and control runs are comparable with the model's internal variability. The HR differences are also very close in magnitude to (and do not exceed) the model's internal variability described in section 4.1.

< Figure 3 >

Let us discuss now prognostic and diagnostic fields such as SST, precipitation, different types of clouds, and time series that are potentially sensitive to changes in the model resulting from using NN emulations. Close similarities have also been obtained for these results of the parallel runs in terms of time mean spatial fields presented in Figs. 3 to 7, which have the same design. The figures contain two columns: the left column shows

results for winter (December-January-February or DJF) and the right column for summer (June-July-August or JJA). The upper raw panels (a) and (d) show fields produced by the full NN radiation NN run, the middle row panels (b) and (e) show mean errors/bias or the difference between the full radiation NN run and the control run (CTL), NN-CTL, and the lower row panels (c) and (f) show the differences between two control runs (i.e., model's internal variability), CTL1-CTL, presented for comparison. Notice that spatial (global) and time mean errors/biases and RSMEs are shown in the panel titles for NN-CTL and CTL1-CTL.

The 17-year (1990-2006) time-mean SST distributions and bias/differences for the full radiation NN run vs. the control run and the differences between two control runs (model's internal variability) are presented for summer and winter in Fig. 3. The SST bias and RMSE for NN-CTL are very small; they are not larger than those of the model's internal variability, CTL1-CTL. The time and global mean errors/biases are near zero and RMSEs are just a small fraction of 1 K. The results for other two seasons (spring and fall) are similar.

< Figure 4 >

Fig. 4 show the 17-year (1990-2006) time-mean distributions and bias/differences for total precipitation (PRATE) for the parallel full radiation NN and control runs for summer and winter, respectively. The PRATE bias is quite limited and occurs mostly in

the tropics; it is also very close in magnitude (as well as RMSE) and pattern to the model's internal variability. The results for other seasons are similar.

Figs.5 to 7 show comparisons for the parallel full radiation NN and control runs for different types of clouds. They present the 17-year (1990-2006) time-mean distributions and bias/differences of total could (Fig. 5), convective precipitation clouds (Fig. 6), and boundary layer clouds (Fig. 7) for summer and for winter. Clouds are very sensitive to any changes in the model and, therefore, provide a suitable and sensitive estimate of the accuracy of NN emulations.

< Figure 5 >

< Figure 6 >

For all types of clouds shown in Figs. 5 through 7, the cloud patterns and bias/differences for parallel total radiation NN and control runs are very close for both seasons presented. The situation is similar for other seasons (spring and fall) and types of clouds such as low, mid, and upper level clouds. The bias is very small and occurs mostly in the tropics. It has the same magnitude (as well as RMSE) and pattern as the differences between two control runs or model's internal variability shown for comparison. For all presented clouds the time and global mean errors/biases are near zero, just ~ 0 - 1%, and RMSEs are just ~1 – 2.5 %.

< Figure 7 >

Let us compare now the results of the parallel NN and control runs in terms of temporal characteristics. The global mean time series for monthly means of the total precipitable water (PWAT), with the seasonal cycle subtracted, are presented in Fig. 8. The figure shows the 17-year (1990-2006) time series for the parallel full radiation NN run (the dash-dotted line) and for two control runs described in Section 4.1 (the solid line for CTL and the dotted line for CTL1). The time series for PWAT presented in Fig. 8 for the parallel full radiation NN and the control run, CTL, show an overall similarity for the entire 17-year (1990-2006) period. The differences between two control runs are similar but marginally larger.

< Figure 8 >

Fig.9 shows the 17-year (1990-2006) time series for the Nino3.4 index for the parallel full radiation NN and the two control runs described in Section 4.1. The Nino3.4 index is calculated over the small area in the equatorial Pacific Ocean shown by the black rectangle in Figs. 3 – 7. The upper panel shows the Nino3.4 index calculated from reanalysis (CDAS), the control runs (the old control – the second panel from the top, and the new control – the second panel from the bottom) and the full radiation NN run (the bottom panel). The time series for the Nino3.4 index are affected by a quite limited SST anomaly sampling for the relatively small area and are very sensitive to any changes in the model or in its computational environment as can be seen from Fig. 9. The explanation for the different details of the Nino3.4 time series is that timing and magnitude of ENSO events is “chaotic” and subject to different phases of internal

variability in the different runs. As can be seen from the standard deviation values included in Fog. 9, the model overestimates the ENSO variability compared with CDAS reanalysis. The overall dissimilarity of the indexes or their deviation from CDAS is not larger than that of the two control runs from CDAS and from each other.

< Figure 9 >

Fig. 10 shows the 17-year (1990-2006) time series for global mean temperature at 850 hPa for the parallel full radiation NN and the two control runs described in Section 4.1. All three time series are close to each other; the differences do not exceed 0.5 K. The small differences between the full radiation NN and control runs are of the same magnitude as those of between two control runs.

< Figure 10 >

The time-mean simulated products presented in Figs. 3-7 as well as other model simulated products show that biases and RMSE for the full radiation NN run are small, i.e., they are overall within the observational errors or uncertainties of reanalysis, and are of a similar magnitude as the model's internal variability.

Close similarity has also been obtained for other model prognostic and diagnostic fields in term of their spatial and temporal characteristics. Summarizing, from the obtained validation results, we can conclude that the differences between decadal climate

simulations produced by the parallel full radiation NN and control runs are overall within or less than the observation errors and uncertainties of reanalysis (e.g. Kalnay et al., 1996). Moreover, these differences (both in terms of bias and RMSE) are of a similar magnitude as the model' internal variability or the differences between two control runs, which are regularly introduced in climate models by routine changes in computer environment (like changes in hardware, operational system, and/or compilers).

4.2.2 Seasonal predictions

We performed similar validation for seasonal predictions for 1990. Basically, the results are similar to those presented above. Fig. 11 shows biases or differences between the NN and control runs (NN-CTL) and differences between two control runs (CTL1-CTL) for seasonal predictions of SST, total clouds (clm CLD), total precipitation (PRATE), and convective clouds (cvl CLD).

All the patterns for the control and NN runs (not shown) are quite close to each other. The differences between seasonal predictions produced by the parallel full radiation NN and control runs are slightly larger than the differences for climate simulations shown above in section 4.2.1. It is partly due to a significantly smaller sample used for seasonal predictions. However, the differences/biases are still comparable with the observation errors and uncertainties of reanalysis. The differences do not increase significantly from season one to season four. For the seasonal predictions, biases or differences between the NN and control runs (NN-CTL) are close by magnitude and do not exceed the differences between the two control runs (CTL1-CTL) or model's internal variability described in

Section 4.1. The time and global mean biases and RMSEs (shown in panel titles) are also quite small for NN-CTL and comparable with those of CTL1-CTL.

< Figure 11 >

The examples of seasonal predictions show overall reasonable results. We realize that for practical implementation seasonal predictions should be produced in an ensemble mode (typically, including several tens of ensemble members), to reduce the impacts of internal variability and to estimate forecast uncertainty. However, this kind of testing is supposed to be done by an implementation group that goes beyond the scope of this study.

5. Conclusions and Discussion

In this study, the NN emulation approach (Krasnopolsky et al. 2005a,b, 2008a) has been further developed and tested in the state-of-the-art, high resolution, coupled NCEP CFS. The developed highly accurate neural network emulations of long-wave (RRTMG-LW and RRTMF-LW) and short-wave (RRTMG-SW) radiation parameterizations are on average 16, 20, and 60 times faster than the original/control long-wave and short-wave radiation parameterizations, respectively (see Table 1). The above results were obtained in a sequential code-by-code comparison and adequately represent the situation when the model is run on a single processor. However, if we compare the control model run using the original parameterizations with the NN run when both runs use multiple processors and threads, the actual speedup will be significantly higher. In the case of parallel calculations utilizing multiple processors and threads, in addition to a significant

speedup, the use of the NN emulations in the model provides an additional advantage, namely it helps to achieve a significantly better load balance.

The use of the full NN model radiation provides: (1) an overall speedup of about 20 – 25% for climate simulations and seasonal predictions, and (2) an opportunity to increase significantly the frequency of radiation calculations (for example, to calculate model radiation at every model dynamic time step) without increasing the total model calculation time.

The full NN model radiation was used for the 17-year climate model simulation and seasonal prediction with the NCEP CFS that has T126 spectral horizontal and high vertical (64 layers) resolutions. We demonstrated the profound similarity for the parallel climate simulations, produced with NN emulations and original radiation (the control run). We have also shown that the model biases and RMSEs associated with NN emulations are quite comparable with internal variability of the model for 17-year integrations and seasonal predictions, which justifies the possibility of using computationally efficient neural network emulations of full model radiation for decadal and longer climate simulations as well as seasonal predictions.

Comparisons with similar results (Krasnopolsky et al. 2008a) obtained for NCAR CAM presented in section 3.6 and Table 1 show that our NN emulation approach works for the high resolution (T126L64) NCEP CFS as well as for the lower resolution (T42L26) NCAR CAM. The NN emulation approach has already been applied to both LWR and

SWR parameterizations and tested in different models with different dynamical cores and with different resolutions (Krasnopolsky and Fox-Rabinovitz 2006a, b, Krasnopolsky et al. 2008a). It is significantly less dependent (in terms of both the accuracy and speedup of calculations) on the increase of vertical resolution than the NN approach introduced by Chevallier et al. (1998) for developing a NN based LWR parameterization NeuroFlux. At vertical resolution of 60 layers and more, both accuracy and rapidity of NeuroFlux can not be achieved simultaneously (Morcrette et al. 2008). As we demonstrated in this study, our NN emulation approach can achieve simultaneously both the desired high accuracy and significant speedup at vertical resolution of 60 layers and more.

Applying the NN emulation approach, which allows us to achieve such a significant speedup with preservation of the accuracy and functional integrity of model physics, may create some challenges that can be resolved using the tremendous flexibility of statistical learning techniques and of the NN technique in particular. Because NN emulations are statistical approximations, there exists a small probability of larger approximation errors or outliers. The major reason for obtaining larger errors is high dimensionality n of the input space of the mapping (1), which reaches several hundreds for NCEP CFS and may reach thousands for future models with significantly higher vertical resolution. It is difficult to sample uniformly a domain in such a high dimensional space. Far corners of the domain may remain underrepresented in the training set. During the NN run, if input vectors belonging to these underrepresented far corners of the domain are encountered, they may cause larger errors in the NN outputs. These larger errors can be successfully controlled using a compound parameterization technique with a quality control procedure

for removing larger errors (Krasnopolsky 2007a, Krasnopolsky et al. 2008b) and/or using the NN ensemble approach with NN emulations (Krasnopolsky 2007b).

The compound parameterization technique can also be used as a method of enriching the training data set by inclusion of underrepresented atmospheric states (Krasnopolsky et al. 2008b). Taking into account the ability of NN to be adjusted sequentially, i.e. using one record of the training set per time, the compound parameterization may be used as a constituent of a dynamically adjustable NN emulation. Such a dynamically adjustable NN emulation will be retrained on-line every time when recognizable changes happen in the environment and an unusual atmospheric state is pinpointed by the compound parameterization. For example, the NN emulations presented in this paper have been trained for the CO_2 level changes in the range of several tens of percents. Dynamically adjustable NN emulations will be able to perform better and under significantly higher changes of the CO_2 level.

Because model vertical resolution determines the NN emulation architecture, i.e., the number of inputs and outputs, every time the vertical resolution of the model is changed (which is usually done quite rarely), the NN emulation needs to be retrained. It is noteworthy that NN retraining can be done routinely and takes a very limited time and effort once the practical framework for a specific model is developed.

In some applications of the developed NN emulation (in a data assimilation system or for an error and sensitivity analysis) not only NN emulation but also its first derivatives (NN

Jacobian) are used. High accuracy of NN emulation does not automatically guarantee the accuracy of the NN Jacobian. An approach that allows us to calculate accurately the NN Jacobian was developed by Krasnopolsky (2007b).

As mentioned above, the NN emulations described in this study have been developed only for the *existing* model parameterizations. Extension of the NN approach to developing new *parameterizations* goes beyond the scope of this study and could be done as a collaborative effort with parameterization developers interested in implementation of more sophisticated and realistic model physics, which are now computationally prohibitive. Also, it is noteworthy that the NN emulation technique can be applied to accelerate calculations of model chemistry and other components.

Acknowledgments. The authors would like to thank Drs. H.-L. Pan, S. Saha, S. Moorthi, and M. Iredell for a valuable help with practical use of NCEP CFS and for useful discussions and consultations. We also thank Drs. S. Moorthi and G. White for reading and commenting on the manuscript. This study is based upon the work supported by the NOAA/CDEP/CTB grant NA06OAR4310047.

MMAB Contribution No. 277

References

Bishop Ch. M., 2006: *Pattern Recognition and Machine Learning*. Springer, 738 pp.

Chevallier, F., F. Chéruy, N. A. Scott, and A. Chédin, 1998: A neural network approach for a fast and accurate computation of longwave radiative budget, *Journal of Applied Meteorology*, **37**, 1385-1397.

_____, J.-J. Morcrette, F. Chéruy, and N. A. Scott, 2000: Use of a neural-network-based longwave radiative transfer scheme in the EMCWF atmospheric model, *Quarterly Journal of Royal Meteorological Society*, **126**, 761-776.

Chou, M.-D., and M. Suarez, 1999: A solar radiation parameterization for atmospheric studies. NASA/TM-1999-104606, Vol. **15**, 40pp.

Clough, S.A., M.W. Shephard, E.J. Mlawer, J.S. Delamere, M.J. Iacono, K.Cady-Pereira, S. Boukabara, and P.D. Brown, 2005: Atmospheric radiative transfermodeling: A summary of the AER codes, *J. Quant. Spectrosc. Radiat. Transfer,* **91**, 233-244, doi:10.1016/j.jqsrt.2004.05.058.

Coˆte´, J., S. Gravel, A. Me´thot, A. Patoine, M. Roch, and A. Staniforth, 1998a: The operational CMC-MRB global environmental multiscale (GEM) model. Part I: Design considerations and formulation, *Mon. Weather Rev*., 126, 1373– 1395.

Coˆte´, J., J.-G. Desmarais, S. Gravel, A. Me´thot, A. Patoine, M. Roch, and A. Staniforth, 1998b: The operational CMC-MRB global environmental multiscale (GEM) model. Part II: Mesoscale results, *Mon. Weather Rev*., 126, 1397–1418.

Hou, Y.-T., S. Moorthi, and K.A. Campana, 2002: Parameterization of solar radiation transfer in the NCEP Models. NCEP/EMC Tech. Mem. 441, 46 pp.

Iacono, M. J., E. J. Mlawer, S. A. Clough, and J.-J. Morcrette, 2000: Impact of an improved longwave radiation model, RRTM, on the energy budget and thermodynamic properties of the NCAR community climate model, CCM3, *J. Geophys. Res.*, 105(D11), 14,873–14,890.

Janiskova´, M., J.-F. Mahfouf, J.-J. Morcrette, & F. Chevallier, 2002: Linearized radiation and cloud schemes in the ECMWF model: development and evaluation. *Quarterly Journal of the Royal Meteorological Society*, 128, 1505–1528.

Kalnay E, et al., 1996: The NCEP/NCAR 40-Year Reanalysis Project. Bull. Amer. Meteorol. Soc., **77**: 437–472.

Kistler R , et al., 2001: The NCEP-NCAR 50-Year Reanalysis, *Bull. Am. Meteorol. Soc.*, 82, 247–268.

Krasnopolsky, V.M., 1997: A neural network forward model for direct assimilation of SSM/I brightness temperatures into atmospheric models, in *Research Activities in Atmospheric and Oceanic Modeling*, CAS/JSC Working Group on Numerical Experimentation, Report No. 25, WMO/TD-No. 792, 1.29-1.30

_____, 2007a: Neural network emulations for complex multidimensional geophysical mappings: applications of neural network techniques to atmospheric and oceanic satellite retrievals and numerical modeling, *Reviews of Geophysi*cs, 45, RG3009, doi:10.1029/2006RG000200.

_____, 2007b: Reducing uncertainties in neural network Jacobians and improving accuracy of neural network emulations with NN ensemble approaches", *Neural Networks,* 20, pp. 454–461

_____, and M.S. Fox-Rabinovitz, 2006a: A new synergetic paradigm in environmental numerical modeling: hybrid models combining deterministic and machine learning components, *Ecological Modelling*, 191, pp. 5–18.

_____, and M.S. Fox-Rabinovitz, 2006b: Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction, *Neural Networks* 19, 122–134.

_____, M.S. Fox-Rabinovitz, and D.V. Chalikov, 2005a: "Fast and accurate neural network approximation of long wave radiation in a climate model", *Monthly Weather Review*, vol. 133, No. 5, pp. 1370-1383.

_____, M. S. Fox-Rabinovitz, and D. V. Chalikov, 2005b, Reply, *Mon. Weather Rev*., 133, 3724–3729.

_____, M.S. Fox-Rabinovitz, A. A. Belochitski, 2008a: "Decadal climate simulations using accurate and fast neural network emulation of full, long- and short wave, radiation.", *Monthly Weather Review,* 136*,* 3683–3695, *doi: 10.1175/2008MWR2385.1*.

_____, M.S. Fox-Rabinovitz, H.L. Tolman, and A. A. Belochitski, 2008b: neural network approach for robust and fast calculation of physical processes in numerical environmental models: compound parameterization with a quality control of larger errors, *Neural Networks*, *doi:10.1016/j.neunet.2007.12.019*

Lacis, A. A., V. Oinas, 1991: A description of the correlated k-distribution method for modeling non-gray gaseous absorption, thermal emission and multiple scattering in vertically inhomogeneous atmospheres. *J.Geophys. Res.* **96**: 9027.9063.

Manners, J., J.-C. Thelen, J. Petch, P. Hill & J.M. Edwards, 2009: Two fast radiative transfer methods to improve the temporal sampling of clouds in NWP and climate models, *Q. J. R. Meteorol. Soc.* **135**: 457 – 468; doi: 10.1002/qj.385

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave, *J. Geophys. Res.*, 102(D14), 16,663–16,682.

Moorthi, S., 2009: Private communications.

Morcrette, J.-J., G. Mozdzynski and M. Leutbecher, 2008: A reduced radiation grid for the ECMWF Integrated Forecasting System, *Monthly Weather Review*, 136, 4760-4772, doi: 10.1175/2008MWR2590.1

_____, P. Bechtold, A. Beljaars, A. Benedetti, A. Bonet, F. Doblas-Reyes, J. Hague, M. Hamrud, J. Haseler, J.W. Kaiser, M. Leutbecher, G. Mozdzynski, M. Razinger, D. Salmond, S. Serrar, M. Suttie, A. Tompkins, A. Untch and A. Weisheimer, 2007: Recent advances in radiation transfer parameterizations, ECMWF Technical Memorandum No. 539, October 18, 2007

Saha, S., S. Nadiga, C. Thiaw, J. Wang, W. Wang, Q. Zhang, H. M. van den Dool, H.-L. Pan, S. Moorthi, D. Behringer, D. Stokes, M. Pena, S. Lord, G. White, W. Ebisuzaki, P. Peng, P. Xie, 2006: The NCEP Climate Forecast System. *Journal of Climate*, Vol. 19, No. 15, pages 3483.3517.

Sato M., J.E. Hansen, M.P. McCormick, and J.B. Pollack, 1993: Stratospheric aerosol optical depths, 1850-1990, *J. Geophys. Res*., 98, 22987-22994.

Venema V., A. Schomburg, F. Ament, C. Simmer: 2007, Two adaptive radiative transfer schemes for numerical weather prediction models. *Atmospheric Chemistry and Physics* **7**(21): 5659 – 5674, URL http://www.atmos-chem-phys.net/7/5659/2007/.

Washington, W.M. and D.L. Williamson (1977), A Description of NCAR GCM's in General Circulation Models of the Atmospheres. *Methods in Computational Physics*, J. Chang Ed., 17, Academic Press, 111-172.

Appendix 1. *Balancing NN radiation outputs*

There exist an integral relationship that relates pressure, heating rates, and fluxes. For example, this relationship for the imbalance, ε, can be written as,

$$\varepsilon = \frac{\sum_{k=1}^{L} \alpha_k \cdot h_k}{\sum_{k=1}^{L} \alpha_k} + \frac{\Phi}{\sum_{k=1}^{L} \alpha_k} = 0. \qquad \text{(A1.1)}$$

$$\Phi = \begin{cases} F_{tup} - F_{\sup} + F_{sdn} & \textit{for LWR} \\ F_{tup} - F_{tdn} - F_{\sup} + F_{sdn} & \textit{for SWR} \end{cases}$$

where $\alpha_k = (p_k - p_{k-1}) / G$, p_k is pressure at a vertical level *k, G* is a constant, F_{tup} is the total sky outgoing LWR or SWR flux at the top of the atmosphere, F_{tdn} is the total sky downward SWR flux at the top of the atmosphere, F_{sup} is the total sky upward LWR or SWR flux at the surface, and F_{sdn} is the total sky downward LWR or SWR flux at the surface.

The outputs of original radiation parameterizations satisfy the relationship (A1.1) with high accuracy because these relationships are explicitly (or implicitly) included into the parameterizations. The outputs of the NN emulations will obviously satisfy (A1.1) only approximately, i.e., in this case the imbalance $\varepsilon \neq 0$; however, ε is small. For example, for the RRTMG – LW NN emulation presented in Table 1, mean value for ε is $6.5 \cdot 10^{-4}$ K/Day. A correction can be introduced for the heating rates, $\tilde{h}_k = h_k + \varepsilon$. The correction makes the corrected or balanced heating rates $\tilde{h}_k$ to satisfy the relationship (A1.1). This correction is very small and, as the results presented in Table 1 show, this balancing procedure does not practically affect the overall accuracy of LWR NN and marginally improve the overall accuracy of SWR NN.

Appendix 2. *Error statistics used in this study*

The statistics presented in Table 1 have been calculated as follows. The mean difference *B* (bias or systematic error of approximation) and the root mean square difference *RMSE* (a root mean square error of approximation) between the original parameterization and its NN emulation describe the accuracy of the NN emulation integrated over the entire 4-D (latitude, longitude, height, and time) data set and are calculated as follows:

$$B = \frac{1}{N \times L} \sum_{i=1}^{N} \sum_{j=1}^{L} [Y(i,j) - Y_{NN}(i,j)]$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \sum_{j=1}^{L} [Y(i,j) - Y_{NN}(i,j)]^2}{N \times L}} \qquad \text{(A2.1)}$$

where $Y(i,j)$ and $Y_{NN}(i,j)$ are outputs from the original parameterization and its NN emulation, respectively, where $i = (latitude, longitude), i=1,\ldots, N$ is the horizontal location of a vertical profile; N is the number of horizontal grid points; and $j = 1,\ldots, L$ is the vertical index where L is the number of the vertical levels.

Using a minor modification of Eq. (A2.1), the bias and RMSE for the mth vertical level of the model can be calculated:

$$B_m = \frac{1}{N} \sum_{i=1}^{N} [Y(i,m) - Y_{NN}(i,m)]$$

$$RMSE_m = \sqrt{\frac{\sum_{i=1}^{N} [Y(i,m) - Y_{NN}(i,m)]^2}{N}} \qquad \text{(A2.2)}$$

The root mean square error has been calculated for each ith vertical profile:

$$prmse(i) = \sqrt{\frac{1}{L} \sum_{j=1}^{L} [Y(i,j) - Y_{NN}(i,j)]^2} \qquad \text{(A2.3)}$$

This error was used to calculate the mean profile root mean square error, *PRMSE*, and its standard deviation, σ_{PRMSE}:

$$PRMSE = \frac{1}{N}\sum_{i=1}^{N} prmse(i)$$
$$\sigma_{PRMSE} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}[prmse(i) - PRMSE]^2} \qquad \text{(A2.4)}$$

The statistics (A2.4) and (A2.1) both describe the accuracy of the NN emulation integrated over the entire 4-D data set. However, because of a different order of integration it reveals different and complementary information about the accuracy of the NN emulations. The root mean square error profile shown in Fig.1 was calculated as:

$$rmsep(j) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[Y(i,j) - Y_{NN}(i,j)]^2} \qquad \text{(A2.5)}$$

Figure captions

Fig.1 Vertical distributions of NN emulation errors for two models: NCAR CAM (26 vertical layers) and NCEP CFS (64 vertical layers). Solid line corresponds to LWR and dashed line to SWR. The errors and their vertical distributions are similar for both models.

Fig. 2 The upper row: zonal and time mean Top of Atmosphere Upward Long (left panel) and Short (right panel) Wave Fluxes (in W per m^2) for the winter. The solid line – the difference (the full radiation NN run – the control (CTL)), the dash line – the differences between two control runs presented for comparison. The lower row: zonal and time annual mean downward (left panel) and upward (right panel) Surface Long Wave Flux (in W/m^2). The fluxes' differences are multiplied by *cos (lat)* to equalize the areas.

Fig. 3 The 17-year (1990-2006) time-mean (NN run) SST distributions and bias/differences for winter (DJF: December-January-February, left column) and for summer (JJA: June-July-August, right column) for the full radiation NN run vs. the control run. The upper row panels show full radiation NN runs. The middle row panels show bias or the difference (full radiation NN run – CTL). The lower row panels show the differences between two control runs shown for comparison. The contour intervals for the SST fields are 5º K and for the SST bias and difference are 0.3º K. Numbers above the figures in the middle and lower rows show the global bias and RMSE.

Fig. 4. The same as in Fig. 3 but for total precipitation (PRATE). The contour levels for the PRATE fields are 2, 4, 8, 16 and 32 mm/day. The contour intervals for the PRATE differences (the bottom panels) are 1 mm/day with 0 mm/day contour skipped for clarity.

Fig. 5 The same as in Fig. 4 but for total clouds. The contour intervals for the cloud fields are 20% and for the differences – 4% with 0 % contour skipped for clarity.

Fig. 6 The same as in Fig. 5 but for convective precipitation clouds. The contour intervals for the cloud fields are 10% and for the differences – 4%.

Fig. 7 The same as in Fig. 5 but for boundary layer clouds. The contour levels for the cloud fields are 10, 20, 40, 60, 80 and 100 % and for the differences – 4%.

Fig. 8 The 17-year (1990-2006) time series of the total precipitable water (PWAT), in kg/m^2, with the seasonal cycle subtracted, for the full radiation NN run (dash-dotted line) and for two control runs described in Section 4.1, CTL (solid line) and CTL1 (dotted line).

Fig. 9 The 17-year (1990-2006) time series for the Nino3.4 index for the reanalysis (CDAS) (the upper panel), and for the parallel full radiation NN (the bottom panel) and two control runs (the middle panels) described in Section 4.1. The Nino3.4 index is calculated over the area in the Pacific Ocean shown by a rectangle in Figs. 3 – 7.

Fig. 10 The 17-year (1990-2006) time series for global mean temperature at 850 hPa (in K) for the parallel full radiation NN (solid line) and the old control (large-dashed line) and new control (short-dashed line) runs.

Fig. 11 Biases or differences between the NN and control runs (NN-CTL) and differences between two control runs (CTL1-CTL) for seasonal predictions of 1990 for: winter (DJF) SST - (a) and (b) panels, summer (JJA) total clouds (clm CLD) – (c) and (d) panels, total precipitation (PRATE) – (e) and (f) panels, and convective clouds (cvl CLD) – (g) and (h) panels. The contour intervals for the SST fields are 1º K, for PRATE – 2 mm/day, for total clouds – 10%, and for convective precipitation clouds – 5%.

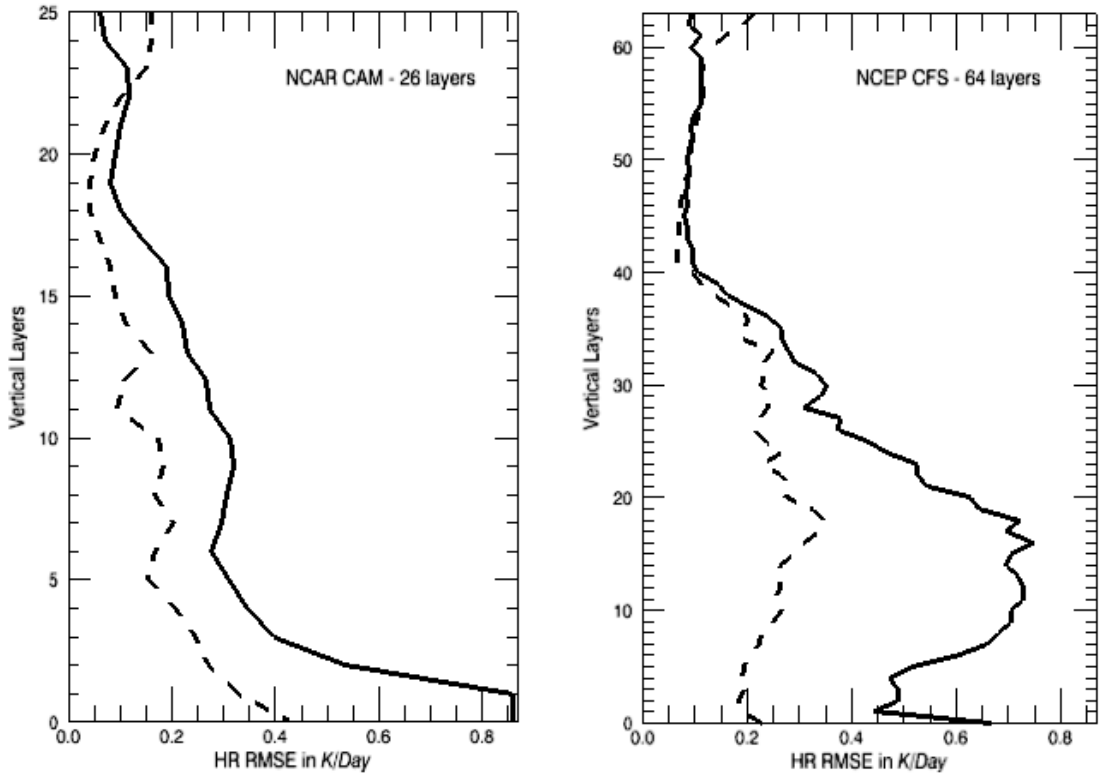

Fig.1 Vertical distributions of NN emulation errors for two models: NCAR CAM (26 vertical layers) and NCEP CFS (64 vertical layers). Solid line corresponds to LWR and dashed line to SWR. The errors and their vertical distributions are similar for both models.

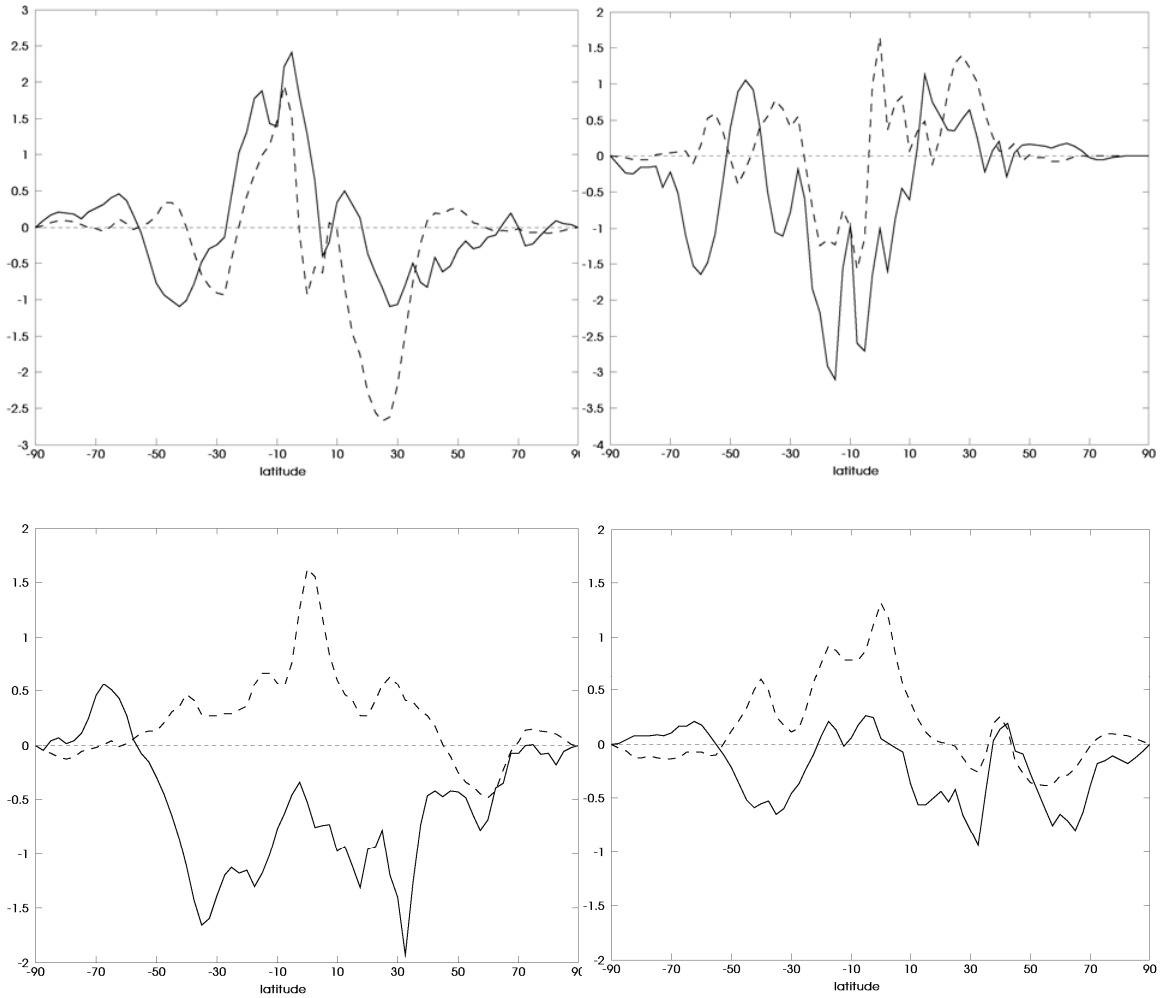

Fig. 2 The upper row: zonal and time mean Top of Atmosphere Upward Long (left panel) and Short (right panel) Wave Fluxes (in W per m^2) for the winter. The solid line – the difference (the full radiation NN run – the control (CTL)), the dash line – the differences between two control runs presented for comparison. The lower row: zonal and time annual mean downward (left panel) and upward (right panel) Surface Long Wave Flux (in W/m^2). The fluxes' differences are multiplied by *cos (lat)* to equalize the areas.

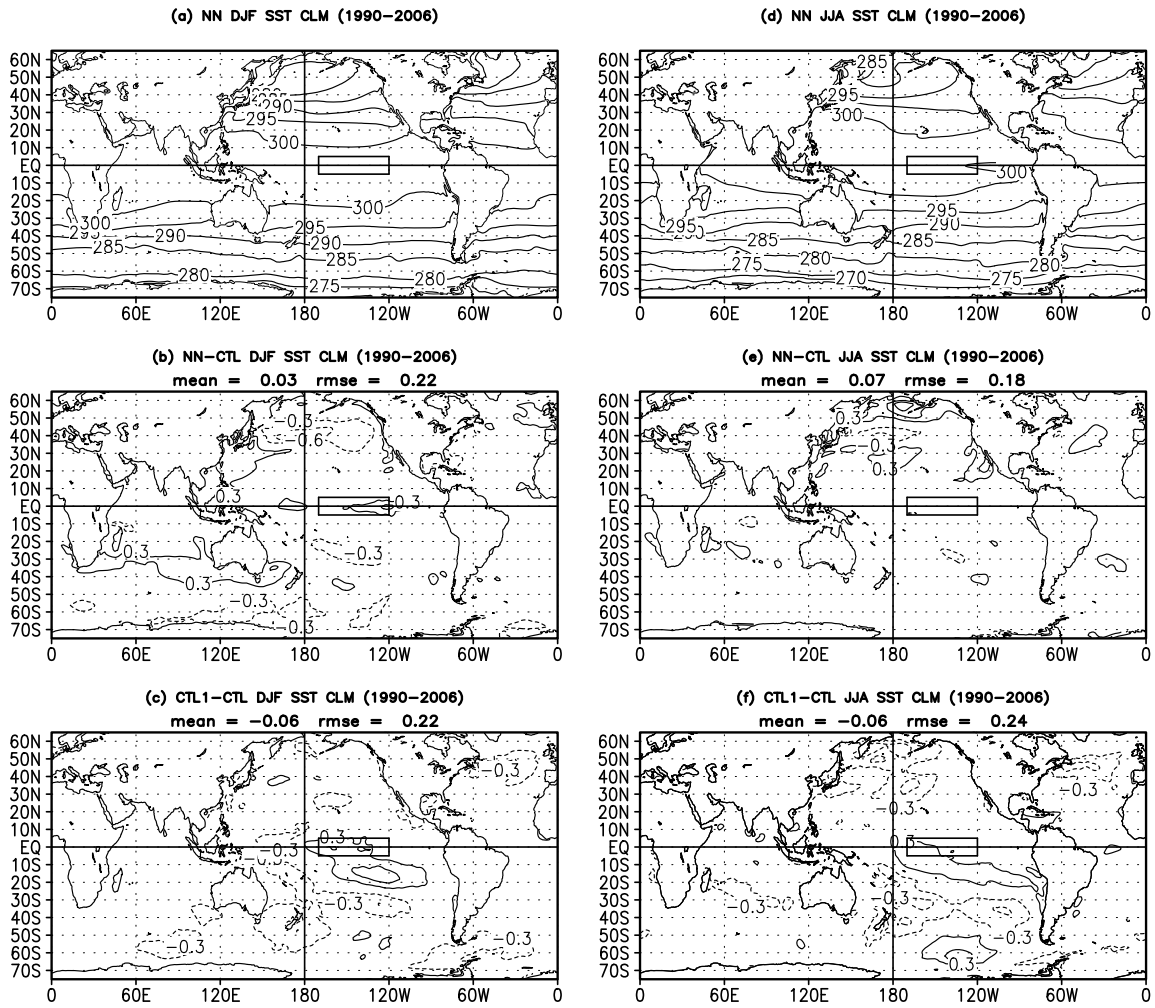

Fig. 3 The 17-year (1990-2006) time-mean (NN run) SST distributions and bias/differences for winter (DJF: December-January-February, left column) and for summer (JJA: June-July-August, right column) for the full radiation NN run vs. the control run. The upper row panels show full radiation NN runs. The middle row panels show bias or the difference (full radiation NN run – CTL). The lower row panels show the differences between two control runs shown for comparison. The contour intervals for the SST fields are 5º K and for the SST bias and difference are 0.3º K. Numbers above the figures in the middle and lower rows show the global bias and RMSE.

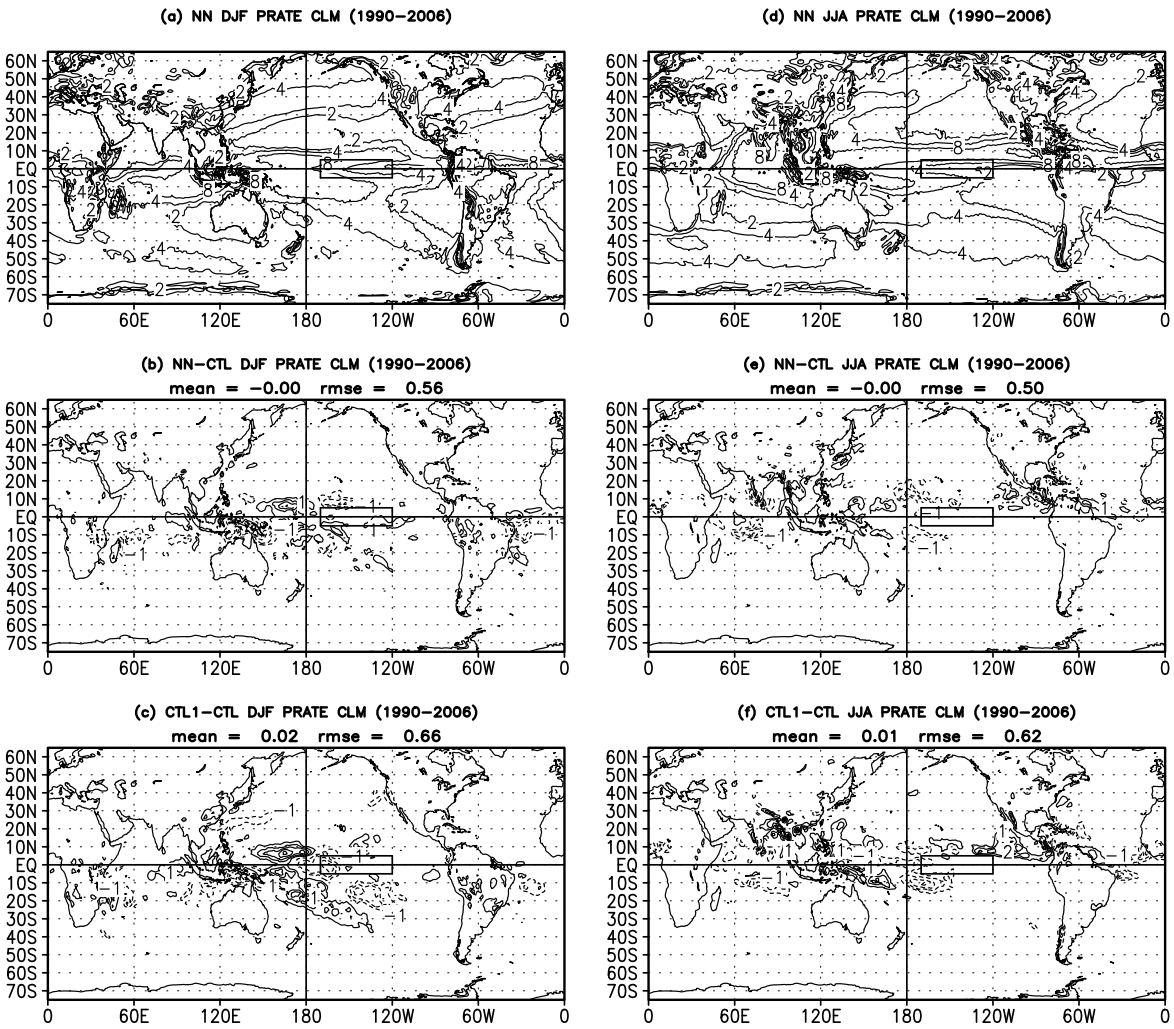

Fig. 4. The same as in Fig. 3 but for total precipitation (PRATE). The contour levels for the PRATE fields are 2, 4, 8, 16 and 32 mm/day. The contour intervals for the PRATE differences (the bottom panels) are 1 mm/day with 0 mm/day contour skipped for clarity.

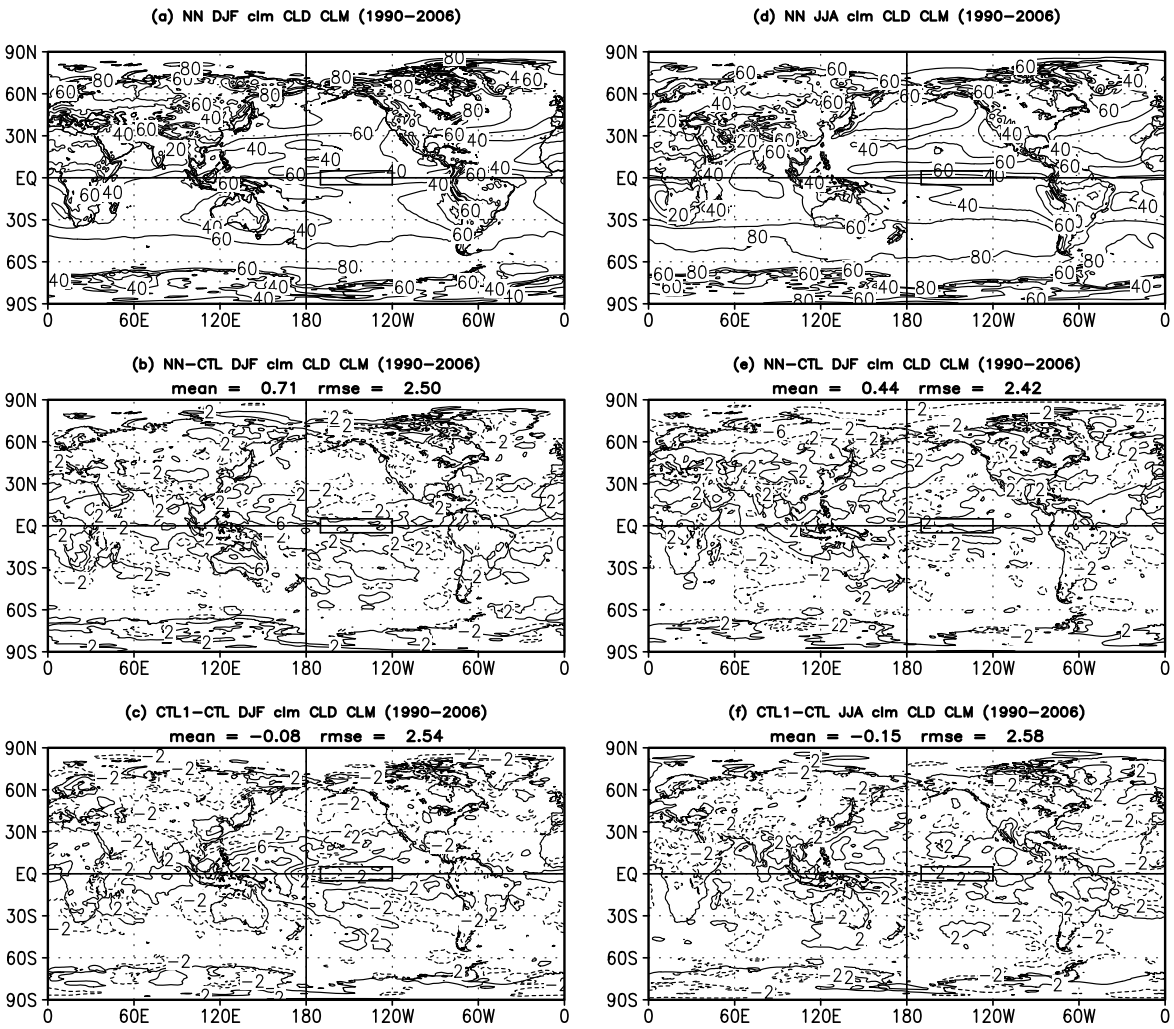

Fig. 5 The same as in Fig. 4 but for total clouds. The contour intervals for the cloud fields are 20% and for the differences – 4% with 0 % contour skipped for clarity.

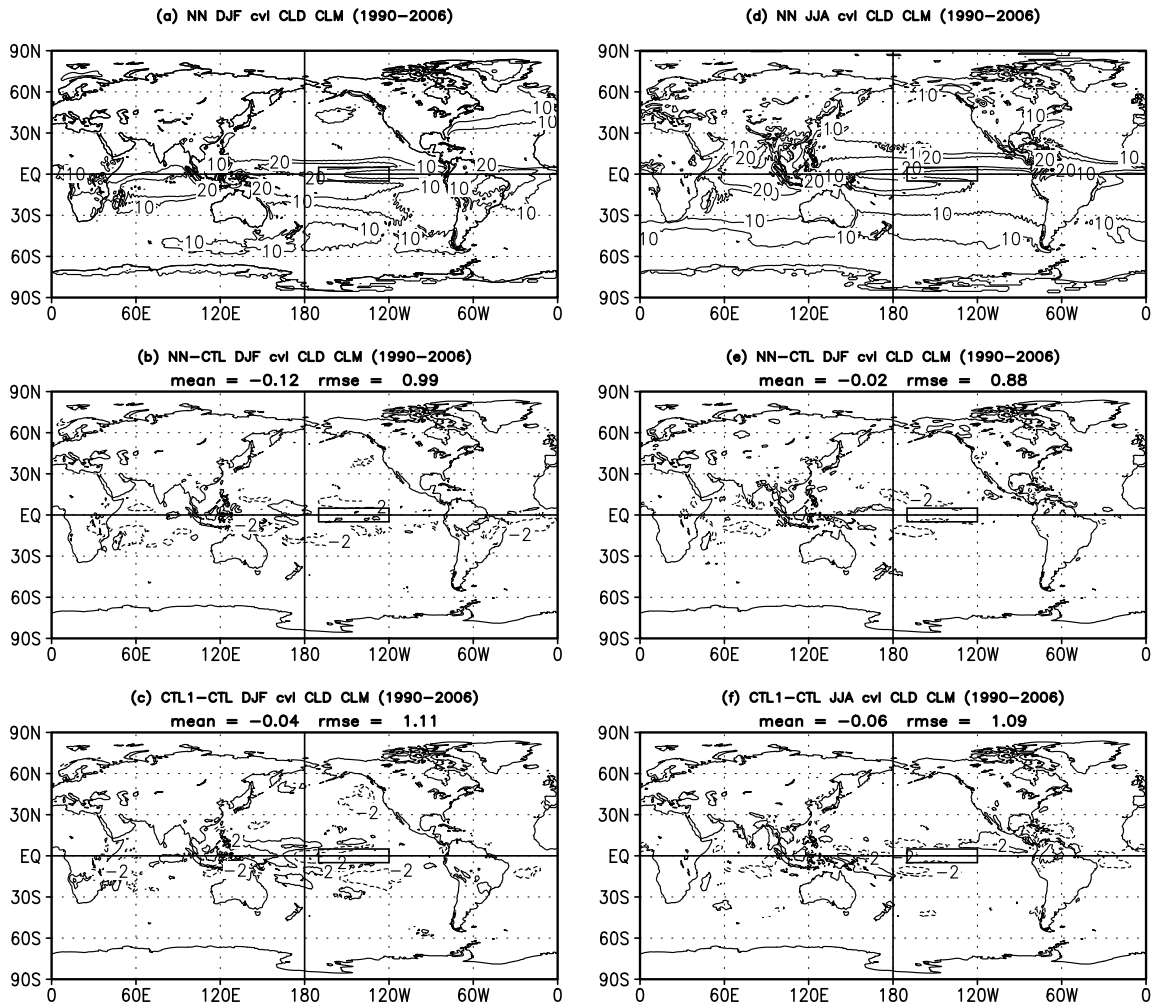

Fig. 6 The same as in Fig. 5 but for convective precipitation clouds. The contour intervals for the cloud fields are 10% and for the differences – 4%.

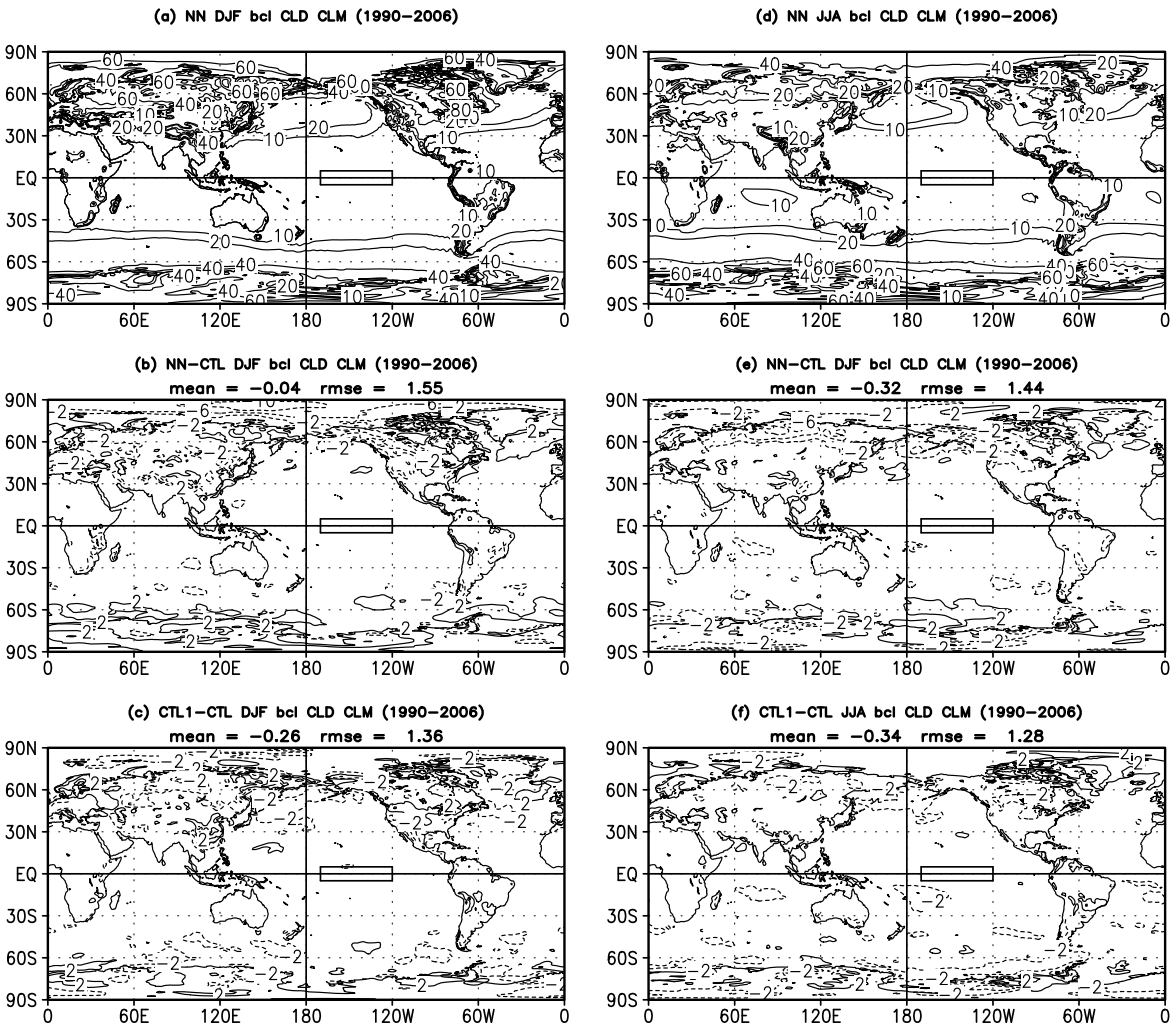

Fig. 7 The same as in Fig. 5 but for boundary layer clouds. The contour levels for the cloud fields are 10, 20, 40, 60 , 80 and 100 % and for the differences – 4%.

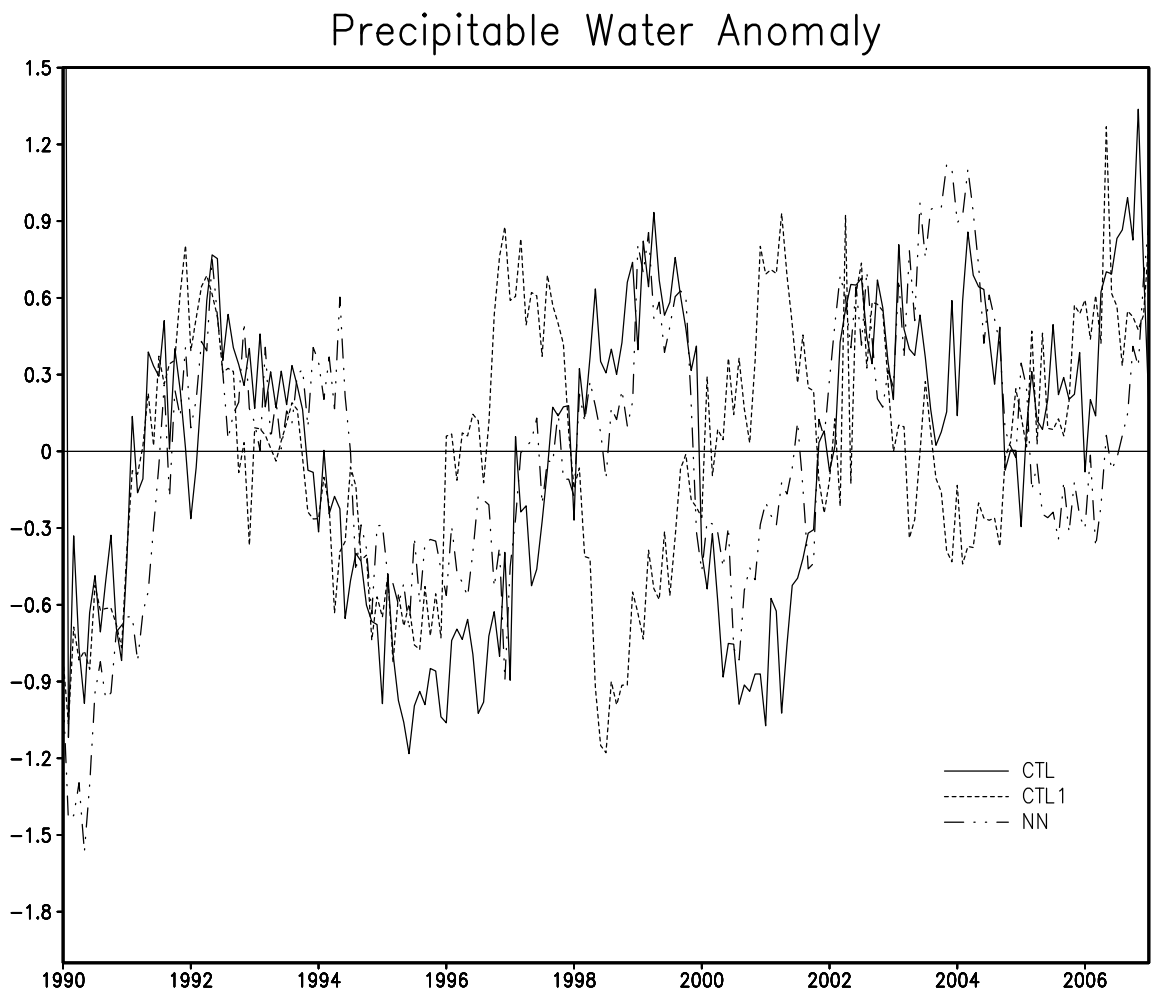

Fig. 8 The 17-year (1990-2006) time series of the total precipitable water (PWAT), in kg/m^2, with the seasonal cycle subtracted, for the full radiation NN run (dash-dotted line) and for two control runs described in Section 4.1, CTL (solid line) and CTL1 (dotted line).

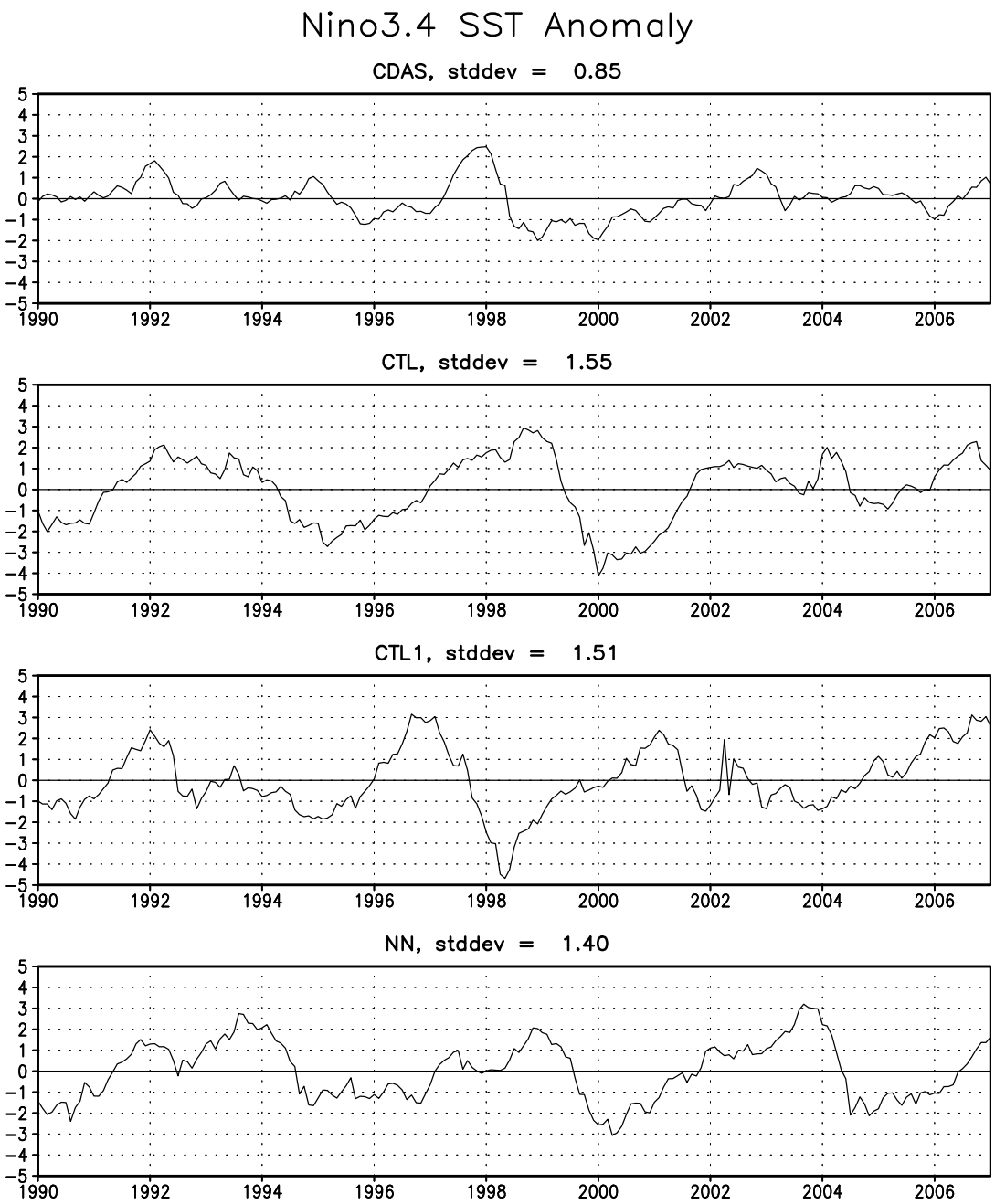

Fig. 9 The 17-year (1990-2006) time series for the Nino3.4 index for the reanalysis (CDAS) (the upper panel), and for the parallel full radiation NN (the bottom panel) and two control runs (the middle panels) described in Section 4.1. The Nino3.4 index is calculated over the area in the Pacific Ocean shown by a rectangle in Figs. 3 – 7.

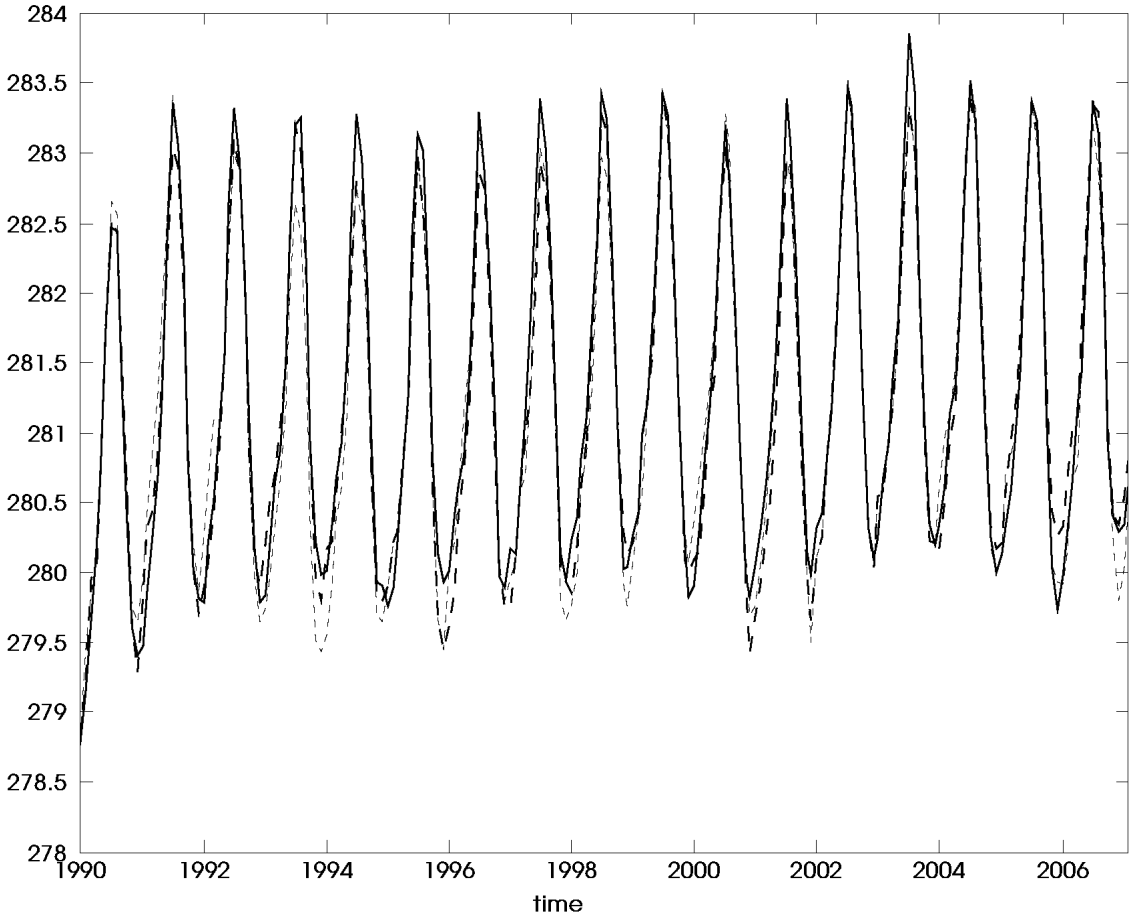

Fig. 10 The 17-year (1990-2006) time series for global mean temperature at 850 hPa (in K) for the parallel full radiation NN (solid line) and the old control (large-dashed line) and new control (short-dashed line) runs.

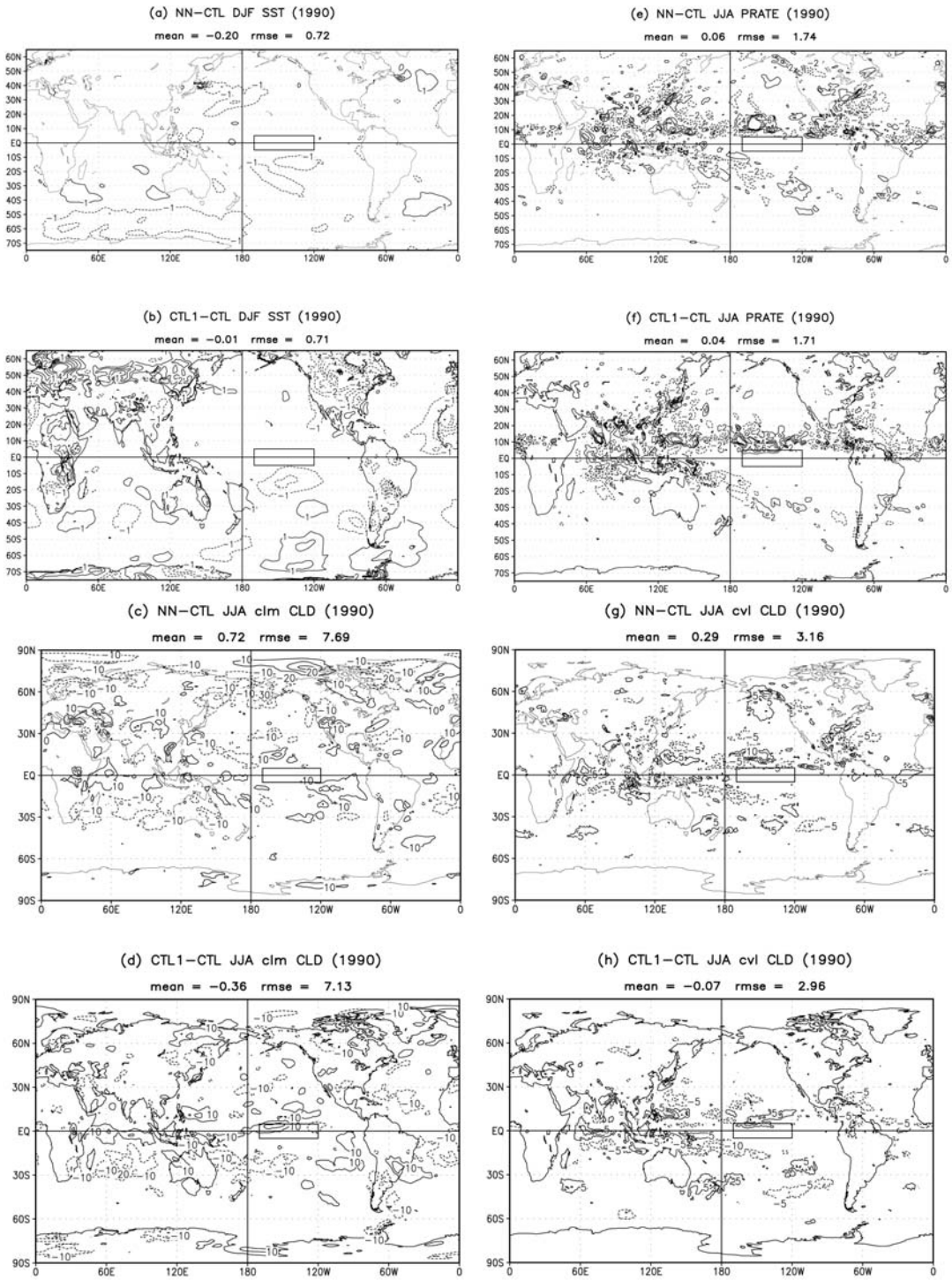

Fig. 11 Biases or differences between the NN and control runs (NN-CTL) and differences between two control runs (CTL1-CTL) for seasonal predictions of 1990 for: winter (DJF) SST - (a) and (b) panels, summer (JJA) total clouds (clm CLD) – (c) and (d) panels, total precipitation (PRATE) – (e) and (f) panels, and convective clouds (cvl CLD) – (g) and (h) panels. The contour intervals for the SST fields are 1º K, for PRATE – 2 mm/day, for total clouds – 10%, and for convective precipitation clouds – 5%.

Table 1. Statistics estimating the accuracy of HRs (in K/day) calculations and the computational performance for NCEP CFS (T126L64) LWR and SWR using NN emulation vs. the original parameterization. For comparison, NCAR CAM (T42L26) LWR and SWR statistics are also shown. Total statistics show the bias, RMSE (A2.1), PRMSE, and σ_{PRMSE} (A2.4) for the entire 3-D HR fields. Layer (for the top and bottom layers) statistics show the bias and RMSE (A2.2) for one horizontal layer (the top or bottom layer). Also, the changes in statistics due to applying the balancing procedure (see Appendix1) are shown for RRTWG[1] LWR and SWR NN emulations. The NN complexity N_C (2) and average speedup[2] η are shown.

Statistics Types	Statistics	LWR				SWR		
		NCAR CAM	NCEP CFS			NCAR CAM	NCEP CFS	
			RRTMG	Change due to Balancing	RRTMF		RRTMG	Change due to Balancing
Total Error Statistics	Bias	$3. \cdot 10^{-4}$	$2.\cdot 10^{-3}$	$6. \cdot 10^{-4}$	$7. \cdot 10^{-4}$	$-4. \cdot 10^{-3}$	$5. \cdot 10^{-3}$	$-3 \cdot 10^{-3}$
	RMSE	0.34	0.49	$1. \cdot 10^{-4}$	0.42	0.19	0.20	$-5. \cdot 10^{-3}$
	PRMSE	0.28	0.39	$3. \cdot 10^{-4}$	0.30	0.15	0.16	$-5. \cdot 10^{-3}$
	σ_{PRMSE}	0.2	0.31	$1. \cdot 10^{-4}$	0.30	0.12	0.12	$1. \cdot 10^{-3}$
Bottom Layer Error Statistics	Bias	$-2. 10^{-3}$	$-1.\cdot 10^{-2}$	$-6. \cdot 10^{-4}$	$6. \cdot 10^{-3}$	$-5. \cdot 10^{-3}$	$9. \cdot 10^{-3}$	$-8. \cdot 10^{-3}$
	RMSE	0.86	0.64	$1. \cdot 10^{-5}$	0.67	0.43	0.22	-0.01
Top Layer Error Statistics	Bias	$-1. \cdot 10^{-3}$	$-9.\cdot 10^{-3}$	$6. \cdot 10^{-4}$	$2. \cdot 10^{-3}$	$2. \cdot 10^{-3}$	$1.3 \cdot 10^{-2}$	$4. \cdot 10^{-3}$
	RMSE	0.06	0.1782	$4. \cdot 10^{-3}$	0.09	0.17	0.21	$1. \cdot 10^{-3}$
NN Complexity	N_C *See eq. (2)*	12,733	33,294	-	93,969	11,418	45,173	-
Speedup, η	*Times*	150	16	-	21	20	60	-

[1] RRTMG and RRTMF are different versions of the radiation code developed by AER Inc. (see Section2 and references there).
[2] Here η shows an averaged (over a global data set) speedup or how many times NN emulation is faster than the original parameterization in a sequential single processor code by code comparison.

Table 2. Comparison of calculation time and speedups $\boldsymbol{\eta}$ for LWR and SWR RRTMG original parameterizations and NN emulations (time is shown per 3,000 profiles) under different cloud conditions. The calculations were performed using a single processor of IBM Power 6 supercomputer.

Parameterization	**LWR RRTMG**			**SWR RRTMG**		
Type of Cloudiness	Clear Sky	3-layer Clouds	Deep Convection	Clear Sky	3-layer Clouds	Deep Convection
Original Parameterization (time in sec)	9.6	10.1	11.7	33.8	42.8	52.9
NN (time in sec)	0.6	0.6	0.6	0.6	0.6	0.6
η (times)	16	16.8	19.5	56	71	88