ISSN 2238-1007

UNCERTAINTIES 2012

Proceedings of the 1st International Symposium on Uncertainty Quantification and Stochastic Modeling
February 26th to March 2nd, 2012, Maresias, São Sebastião, SP, Brazil

GLOBAL DATA ASSIMILATION USING ARTIFICIAL NEURAL NETWORKS IN SPEEDY MODEL

Rosangela. S. Cintra[1], **Haroldo F. de Campos Velho**[1]

[1] Brazilian National Institute of Space Research
São José dos Campos, SP, Brazil
e-mail: rosangela.cintra@lac.inpe.br, haroldo@lac.inpe.br

Abstract. *Weather forecasting systems require a model for the time evolution and an estimate of the current state of the system. Data assimilation provides such an initial estimate of the atmosphere where it combines information from observations and from a prior short-term forecast producing an current state estimate. An Artificial Neural Network (ANN) is designed for data assimilation. The use of observations from the earth-orbiting satellites in operational numerical prediction models is performed for improving weather forecasts. The data related to atmospheric, oceanic, and land surface state from satellites provides increasingly large volumes. However, the use of this amount of data increases the computational effort. The goal here is to simulate the process for assimilating temperature data computed from satellite radiances. The numerical experiment is carried out with the global model Simplified parameterizations, primitive-Equation Dynamics (SPEEDY) with simplified physical processes of an atmospheric general circulation in tri-dimensional coordinates. For the data assimilation scheme was applied an ANN: a Multilayer Perceptron(MLP) with supervised training. The MLP-ANN is able to emulate the analysis from the Local Ensemble Transform Kalman Filter(LETKF). LETKF is a version of Kalman Filter with Monte-Carlo ensembles of short-term forecasts. In this experiment, the MLP-ANN was trained with supervision from first six months considering the years 1982, 1983, and 1984. A hindcasting experiment for data assimilation performed a cycle for january of 1985 with MLP-NN, LETKF and SPEEDY model. The synthetic temperature observations were used. The numerical results demonstrate the effectiveness of this ANN technique on atmospheric data assimilation. The results for analysis with ANN are very close with the results from LETKF data assimilation. The simulations show that the major advantage of using MLP-NN is the better computational performance, with similar quality of analysis. The CPU-time assimilation with MLP-NN is 75% less than LETKF with the same observations. Actually, considering the supervised ANN for data assimilation, the most relevant issue is the computational speed-up for computing the analyzed initial condition for state model that accelerates the whole process of numerical weather prediction.*

Keywords. *data assimilation, artificial neural network, ensemble kalman Filter, numerical weather forecasting.*

1 INTRODUCTION

The procedure which takes atmospheric observed data and creates meteorological fields over some spatial or temporal domain is usually called analysis or data assimilation, when the data are distributed in time and the procedure uses an explicit dynamical model for the time evolution of the atmospheric flow. The fields produced by an analysis or an assimilation must satisfy two basic requirements. On the one hand, they must be close to the observations, at the required spatial and temporal scales. On the other hand, they must verify dynamical and/or statistical relationships which are known to be satisfied by the real atmospheric fields. The numerical weather prediction (NWP) was confronted with having to solve an initial-value problem. Hence, inverse methods were carried over from solid-earth geophysics to estimate, at first, the state of an idealized atmospheric steady-state field. Data assimilation is the objective melding of observed information with model-predicted information. Data assimilation rigorously combines statistical modeling with physical modeling; thus, formally connecting the two approaches. (Daley,1991) is the standard text on data assimilation.

The observational data used in data assimilation are conventional data and satellite data. These data include surface observations and balloon soundings, as well as ship and aircraft observations. Operational satellite data are taken and processed in real-time and distributed around the world. Though small in number, in meteorological data assimilation the conventional data are very important to the quality of the analysis and the forecast. The satellite data assures high quality global analyses. It is very clear that assimilation of satellite observations will make a key contribution to that improvement in forecast skills, given the future growth (five orders of magnitude increase in satellite data over ten years) and improvement of the global observing system expected in the area of space-borne observing systems. As a result there is a need an assimilation method able to get the initial field for the numerical model in time to make a prediction. At present most NWP centers cannot assimilate all the data due to computational costs and limitations in storing the data.

Operational satellite data are taken and processed in real-time and distributed around the world. In the case of satellite observations the measurements are radiances and the observation operator (in data assimilation scheme) might include a forward radioactive transfer calculation from the model"s geophysical parameters to radiance space. The analysis is the best estimate of the state of the system based on the optimization criteria and error estimates. The computational challenge to the traditional techniques of data assimilation lies in the size of matrices involved in operational NWP models, currently running at a million equations (equivalent to full matrix elements of the order of 10^{12}!). In this scenario the applications

ISSN 2238-1007

of ANN in data assimilation were suggested. The ANN technique uses neural networks to implement the function: $x^a = F_{rna}[y^o, x^o, (x^{trn})]$, where F is the data assimilation process, x^a is the analysis field with innovation that represents the observation-based correction to the model; y^o are observations of the constituent, x^f is a model forecast, simulated, estimates of the constituent often called the first guess and x^{trn} is the analysis field only for training ANN.

Methods using Artificial Neural Networks (ANN) have been proposed showing consistent results regarding implementation in simple models, see (Nowosad, 2001; Harter, 2004; Furtado, 2008; Cintra, 2010). This paper presents an experiment using an Atmospheric General Circulation Model (AGCM): model SPEEDY (*Simplified Parameterizations PrimitivE-Equation Dynamics*), which is a 3D dynamic model, with simplified physics parameterization (Molteni,2003). The SPEEDY model is close to the AGCM used in the operational centers. A set of Multilayer Perceptron (MLP) [Haykin] are employed, with training to emulate the analysis produced by the LETKF (*Local Ensemble Transform Kalman Filter*) (Hunt,2004; Bishop,2001).

The goal here is to simulate the process for assimilating temperature data computed from satellite radiances, introducing a new methology for weather forecasting. The performance would be faster than conventional schemes for data assimilation, with the same quality of the emulated analysis. The tests are performed by using the SPEEDY model.

2 METHODOLOGY

The experiment was conducted with the forecast from the SPEEDY model, employing the analysis obatianed by the LETKF. The analysis is emulated by a MLP-ANN.

2.1 The SPEEDY Model

SPEEDY is an atmospheric general circulation model (AGCM) developing to study on global-scale dynamics and numerical weather prediction. The dynamic variables on the primitive meteorological equations are integrated by spectral method in the horizontal at each vertical level (see: Bourke, 1974; Held and Suarez, 1994; Myioshi, 2005). The model has a simplified set of physical parameterization schemes, but they are similar to realistic weather forecasting numerical models.

The model configuration used in this paper is a global model with spectral resolution of **T30L7** (horizontal truncation of 30 wave components in the expansio, and seven vertical levels), corresponding to regular grid with 96 zonal points (longitude), 48 southern points (latitude), and 7 vertical pressure levels ($100, 200, 300, 500, 700, 850, 925$ hPa).

According to Molteni (2003), the SPEEDY model simulates the general structure of global atmospheric circulation fairly well, and some aspects of the systematic errors are similar to many AGCMs. The boundary conditions of the SPEEDY model includes topographic height and land-sea mask. The SPEEDY model is a hydrostatic model in σ-coordinates, and the transformed vorticity-divergence scheme is described by Bourke (1974). The prognostic variables of input and output model are the absolute temperature (**T**), surface pressure (**ps**), component of zonal wind (**u**), component of southern wind (**v**), and specific humidity (**q**).

2.2 LETKF

The LETKF was proposed by Hunt and co-authors (2004) as an efficient upgrade of LEKF (Ott et al., 2004). The Bayesian approach can be identified as ensemble Kalman Filter (EnKF) and particle filtering (PF) methods, two efficient and flexible Monte-Carlo methods to solve the optimal filtering problem. EnKF is a sequential filter method, which means that the model is integrated forward in time and, whenever observations are available, the boservations are used to reinitialize the model, before the integration continues. The LETKF separate the entire global grid into independent local patches.

The basic idea of LETKF is to perform the analysis at each grid point simultaneously, using the state variables and all observations in the local region centered at that point. Each member of the ensemble gets its forecast: $x^f_{n,i}$ $(i = 1, 2, \cdots, K)$, where $K$is the total members at time t_n. For estimating the state vector of the reference model is used the mean of the ensemble forecasts:

$$\bar{x}^f_n = \frac{1}{K}\sum_{i=1}^{K} x^f_{n,1} \tag{1}$$

and the model error covariance matrix is computed as

$$\mathbf{P^f} = \frac{1}{K-1}\sum_{i=1}^{K}(x^f_{n,1} - \bar{x}^f_n)(x^f_{n,1} - \bar{x}^f_n)^T \ . \tag{2}$$

LETKF in the local analysis allows different linear combinations of the ensemble members for different regions, and the comprehensive analysis explores a larger spatial scale. For local implementation separates groups of neighboring observations at a central point for a region of the model grid. Each grid point has on local patch, the number of vector components is equal to the number of global grid points, each local patch is treated independently (Miyoshi,2005).

2.3 Artificial Neural Netwoks (ANN)

ANN is computational system with massively parallel and distributed processing. Multilayer Perceptron (MLP) is an ANN architecture, where the interconnections from inputs to the output layer has *at least* one intermediate layer of

ISSN 2238-1007

neurons, called hidden layer, it has also *at least*, one input vector: $\mathbf{x} = [x_0\ x_1\ x_2\ \cdots\ x_N$, and one output vector $\mathbf{s} = [s_0\ s_1\ s_2\ \cdots\ s_N$. A Multilayer network performs a complex mapping $\mathbf{s} = \Psi(\mathbf{w},\mathbf{x})$ parameterized by the synaptic weights $\mathbf{w}$. The set of well-defined procedures to adjust the weights $\mathbf{w}$ of an ANN is applied to produce a desired output. Such procedures are also called a learning (or training) algorithm. The *backpropagation* scheme is one of the most known procedure for training a MLP. This algorithm is a supervised process, where the network receives input vectors with their corresponding response or desired output. MLP with backpropagation learning algorithm, commonly referred to as backpropagation neural networks. These are feedforward networks, whose aim is to extract high order statistics from the input data (Miki, 1999). Figure 1 depicts a backpropagation neural network with a hidden layer.

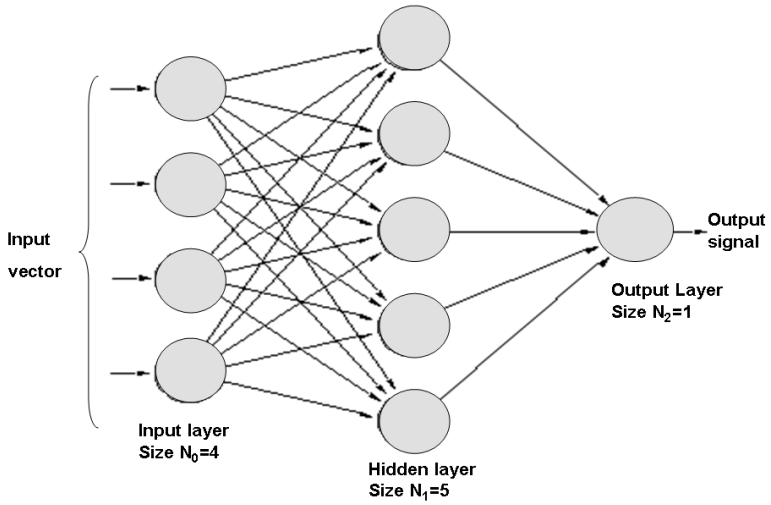


Figure 1: Multilayer Perceptron example.

The training process determines the synaptic weights when minimize the error between the output calculated by the network and the expected response to some input vectors.

Neural networks will solve nonlinear problems, if nonlinear activation functions are used for the hidden and/or the output layers. Several activation functions can be used, but the sigmoid functions are commonly used:

$$\text{logistic function}: \qquad \varphi(v) = \frac{1}{1+\exp(-av)};$$

$$\text{tangent hiperbolic function}: \qquad \varphi(v) = \frac{1-\exp(-av)}{1+\exp(-av)}\ . \tag{3}$$

After learning, the set of synaptic weights is able to activate the neurons of the MLP and get results for entries that were outside the set of training data, generalizing the information learned.

2.4 Experimental settings

The MLP designed here has *two* (2) inputs (model variables and forecast vectors), one (1) neuron in the output (to analysis vector), eleven neurons in a hidden layer, the activation function to ensure nonlinearity of the problem used was the tangent hyperbolic function.

The observational data used in data assimilation are synthetic satellite data. These observations are radiance. From a inversion procedure, it is possible to compute temperature profile from radiances. The satellite data have been essential of weather forecasts. The observations were generated from ”true” model fields adding a random noise. The variables were palced at some grid point model. The grid points chosen for simulating satellite observations, obtaining values for a merged point model (alternating: a grid point has observation, and another grid point has no observation). Both assimilation schemes, LETKF or ANN, use the same numbers of observations.

In this configuration, the SPEEDY model was run for a long time integrations of the state, to create ”true” values to start the integrations of the model. The true integration of the model was made for three years: from January 01 (1982) up to January 31 (1985), generating outputs in four times a day (00, 06, 12 and 18 UTC). The LETKF was performed with those synthetic observations of the temperature to generate the vector analysis and obtaining the desired output to train the neural network. The executions of the model with LETKF were made for the cycle of 6 hours, as mentioned before.

For the ANN data assimilation scheme, we define a local observation *influence* operator, on the neighboring from the grid points. This operator was emloyed for the zonal and southern winds, for the vertical boundaries (at the bottom and

ISSN 2238-1007

top levels). This calculation was based on the distance from its neighbor:

$$\hat{y}^o = \frac{1}{r^2}\left[y_l^o + \sum_{l=1}^{N}(y_l^o + \delta)\right] \tag{4}$$

$$r = \mathrm{dist}(y_l^o + \delta,\ y_l^o) \tag{5}$$

where N is the total neighbors grid points *without* observations, δ is a cubelike shape characterized by the horizontal and vertical grid lengths, this is based on the distance of each observation point:

$$\delta = (x_i^f - y_i^o)^2 + (x_j^f - y_j^o)^2 + (x_k^f - y_k^o)^2$$

where x^f is grid point with forecst value but without observation, y^o is observation in a grid point; the subscripts i, j, k, are the indices to identify the coordinates: latitude, longitude, and vertical level, respectively.

Another strategy used to accelerate the processing of MLP training was to divide the entire globe into six regions. Each hemisphere of the planet (North and South) were divided into three regions. This division is based on size of regions, but the number of radiosondes observations are distinct, illustrate by Figure 2. The division and influence observations were applied in LETKF program before its performance, because the input data for ANN was already separated in training areas. However, the LETKF procedures are not modified.

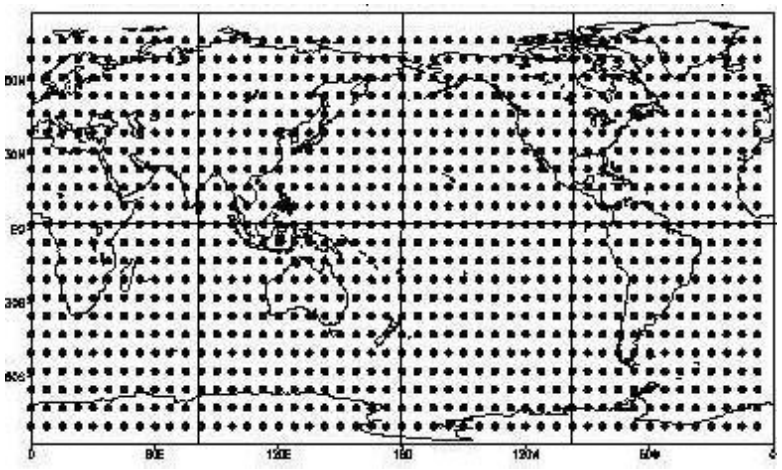


Figure 2: Observations localizations in global area. The dot points represent satellite (about 1056) divided in six regions.

There are 6 different MLP-NN, one for each region.

The network was trained to emulate the LETKF analysis. The training was made with collected data of the first six months of years 1982, 1983, and 1984. The MLP generalization is initiated with ANN data assimilation one cycle in the first on January 01 (1985) at 00 UTC, generating predictions for the SPEEDY model with new global analysis, until January 10 (1985) at 18 UTC. The ANN was re-trained (with true model field), and re-start another cycle at January 11 (1985) at 00 UTC, until January 20 (1985) at 18 UTC. After the second cycle, the ANN was retraining again, and the trained MLP-NN was applied to the last days to the January 1985.

3 RESULTS

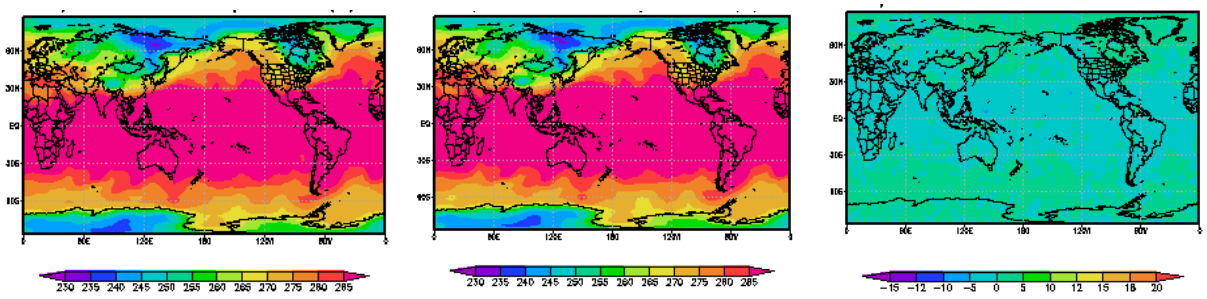
The input and output values were processed on grid points for time integrations, alternatimg forecasting and analysis cycle to assimilate the temperature (T). The results show the analysis fields generated by the MLP-ANN and by LETKF data assimilation schemes: for 03/Jan/1985 at 12 UTC, showing two levels: 950 hPa and 500 hPa (see Figures 3 and 4). The analysis for the day 17/Jan/1985 at 950 hPa is shown in Figure 5, and for the day 22/Jan/1985 at 500 hPa is presented in Figure 6. The figures present global LETKF and MLP-ANN analysis fields, and the differences between the ”true” field and observations, and the MLP-ANN analysis field.

4 CONCLUSION

The results show that the application of MLP-NN as assimilation system produces an analysis similar to the LETKF assimilation system. The first conclusion of this experiment is: the MLP-ANN can emulate the LETKF analysis for the temperature, with observations provide from satellites sensors. The computational performance of the MLP-NN is better than LETKF.

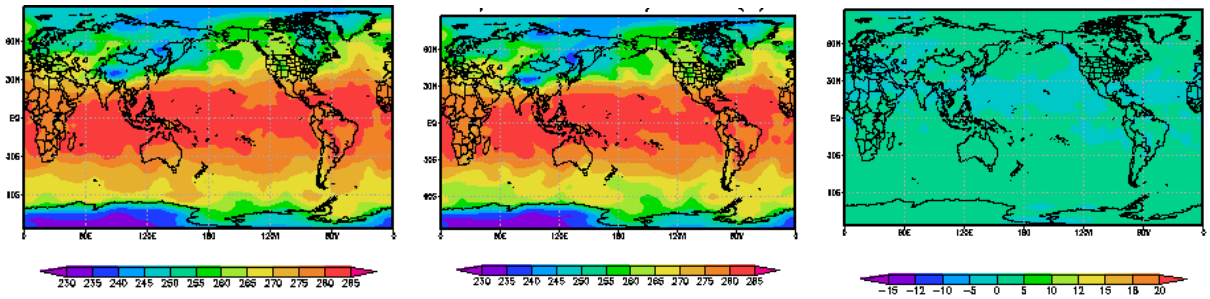
There are several aspects of the modeling and assimilation problem that stress computational systems and push capability requirements. The common ones in modeling are increased resolution, improved physics, inclusion of new processes, and integration and concurrent execution of Earth-system components. Often, real-time needs define capability requirements. When considering data assimilation, the computational requirements become much more challenging. The use

ISSN 2238-1007



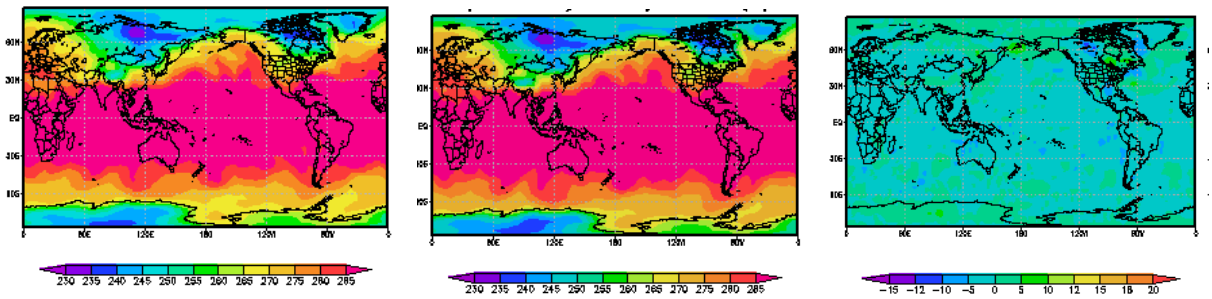
(a) ANN Analysis SPEEDY Field (b) LETKF Analysis SPEEDY Field (c) ANN-LETKF

Figure 3: Analysis fields for Temperature to 03/01/1985 12 UTC to level 950 hPa.



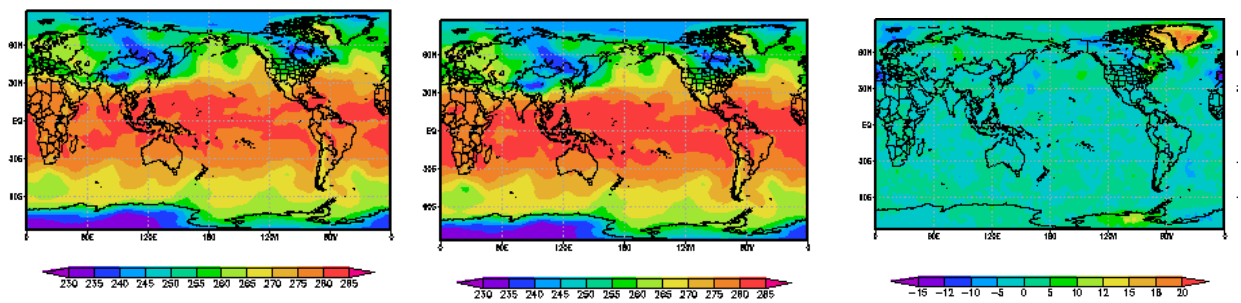
(a) ANN Analysis SPEEDY Field (b) LETKF Analysis SPEEDY Field (c) ANN-LETKF

Figure 4: Analysis fields for Temperature to 03/01/1985 12 UTC to level 500 hPa.



(a) ANN Analysis SPEEDY Field (b) LETKF Analysis SPEEDY Field (c) ANN-LETKF

Figure 5: Analysis fields for Temperature to 17/01/1985 12 UTC to level 950 hPa.



(a) ANN Analysis SPEEDY Field (b) LETKF Analysis SPEEDY Field (c) ANN-LETKF

Figure 6: Analysis fields for Temperature to 22/01/1985 12 UTC to level 500 hPa.

ISSN 2238-1007

of observations from the earth-orbiting satellites in operational numerical prediction models is performed for improving weather forecasts. However, the use of this amount of data increases the computational effort. As a result, it is important to investigate schemes to deal with huge amount of data, looking at the period of time available to make a prediction. At present, most numerical weather prediction centers to assimilate all the data use some approach to select the data to be assimilated, in order to reduce the computational costs.

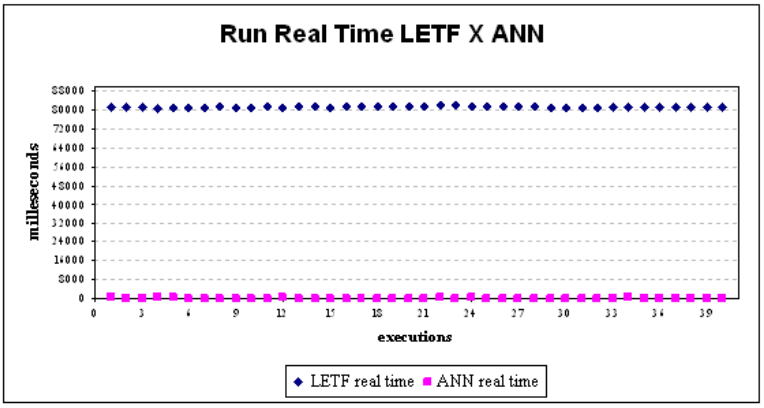


Figure 7: Computational performance for 10days for data assimilation cycle (before retraining).

The Figure 7 shows a cycle of 40 insertions in the data assimilation process: ten (10) days with 7392 satellite observations, run four times per day. The time was measured in milliseconds, for both data assimilation schemes. The ANN computational performance was higher than the performance of the LETKF system, i.e. these results show that the computational efficiency of neural network to the problem of atmospheric data assimilation is better with the similar quality in analysis field. The CPU-time assimilation with MLP-NN is 80 times faster than LETKF, in our numerical experiment.

Table 1: CPU-run time of 124 cycles of data assimilation (analysis and forecasting).

CPU-time of 124 cycles	MLP-NN (hh:mm:ss)	LETKF (hh:mm:ss)
Total run time	**00:04:11**	**04:47:43**
Total real Analysis time	**00:00:22**	**02:45:50**

Table 2: CPU-run time of 40 cycles of data assimilation (analysis and forecasting).

CPU-time of 40 cycles	MLP-NN (hh:mm:ss)	LETKF (hh:mm:ss)
Total real run time	**00:00:09**	**00:54:15**

The performance of MLP-NN for data assimilation is capable to increase the number of observations, supporting the research with global models to test different resolutions and/or more observations. Actually, considering the supervised ANN for data assimilation, the most relevant issue is the computational speed-up for computing the analyzed initial condition, see Figure 7 and Tables 1 and 2.

5 ACKNOWLEDGEMENTS

The authors want to thank Dr. Takemasa Miyoshi and Prof. Dr. Eugenia Kalnay for providing routines for the SPEEDY model and the LETKF system.

6 REFERENCES

Bishop, H. C., Etherton. B. J., Majumdar, S. J., 2001, “Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I. Theoretical Aspects”. Monthly Weather Review, Vol. 129, pp. 420–436.

Bourke, W., 1974, “A multilevel spectral model: formulation and hemispheric integrations.” Mon. Wea. Rev., Vol. 102, pp. 687–701.

Cintra, R. S., 2010, “Assimilação de Dados com Redes Neurais Artificiais em Modelo de Circulação Geral da Atmosfera”. D.Sc. dissertation on Applied Computing, National Institute for Space Research (INPE: Instituto Nacional de Pesquisas Espaciais), São José dos Campos, Brazil.

ISSN 2238-1007

Daley, R., 1991, "Atmospheric data analysis". Cambridge University Press.

Furtado, H. C., 2008, "Redes neurais e diferentes métodos de assimilação de dados em dinâmica não linear". M.Sc. thesis on Applied Computing, National Institute for Space Research (INPE: Instituto Nacional de Pesquisas Espaciais), São José dos Campos, Brazil.

Gardner, M.W.; Dorling, S. R., 1998, "Artificial neural networks: the multilayer perpectron: Review of Applic. in the Atmospheric Sciences". Atmospheric Environment, Vol. 32 (14/15), pp. 2627–2636.

Härter, F.P., 2004, "Redes Neurais Recorrentes Aplicadas á Assimilção de Dados em Dinâmica Não Linear". D.Sc. dissertation on Applied Computing, National Institute for Space Research (INPE: Instituto Nacional de Pesquisas Espaciais), São José dos Campos, Brazil.

Haykin, S., 2001. "Redes neurais - princípios e práticas ". Bookmann, Porto Alegre.

Hunt, B., Kalnay, E., Kostelich, E. J., Ott, E., Patil, D., Sauer, T., Szunyogh, I., Yorke, J. A., Zimin, A. V., 2004, "Four–dimensional ensemble kalman filtering". Tellus, Vol. 56A, pp. 273–277.

Kalman, R. E., 1960, "A new approach to linear filtering and prediction problems". Trans. of the ASME–Journal of Basic Engineering, Vol. 82(D), pp. 35–45.

Held, M.L.; Suarez, L., 1994, "A proposal for the intercomparison of dynamical cores of atmospheric general circulation models". Bull. Am. Meteorol. Soc., Vol. 75, pp. 1825–1830.

Miki, F.T., Issamoto, E., da Luz, J.I., de Oliveira, P.B. de Campos Velho,, H. F., da Silva, J.D., 1999, "A Neural Network Approach in a Backward Heat Conduction Problem", Brazilian Conference of Neural Networks, São José dos Campos (SP), Brazil.

Miyoshi, T., 2005, "Ensemble Kalman Filter experiments with a primitive–equation global model". PhD thesis (on Atmospheric Science). University of Maryland, College Park, Maryland, USA.

Molteni, F., 2003, "Atmospheric simulations using AGCM with simplified physical parameterizations. I: model climatology and variability in multi-decadal experiments". Clim. Dyn., Vol. 20, pp. 175–191.

Nowosad,A. G., 2001, "Novas Abordagens para Assimilação de Dados Meteorológicos", D.Sc. dissertation on Applied Computing, National Institute for Space Research (INPE: Instituto Nacional de Pesquisas Espaciais), São José dos Campos, Brazil.

Ott, B.; Hunt,,B; SZUNYOGH, I.; Zimin, A. V. ; Kostelich, E. J. ; Corazza, M.; Kalnay, E.; Patil, D. J.; Yorke, J. A., 2004, "A local ensemble Kalman filter for atmospheric data assimilation". Tellus, vol. 56(a), pp. 415–428.

RESPONSIBILITY NOTICE

The authors are the only responsible for the printed material included in this paper.