Neural network uncertainty assessment using Bayesian statistics with application to remote sensing: 2. Output errors

F. Aires

Department of Applied Physics and Applied Mathematics, Columbia University/NASA Goddard Institute for Space Studies, New York, USA

CNRS/IPSL/Laboratoire de Météorologie Dynamique, École Polytechnique, Palaiseau, France

C. Prigent

CNRS, LERMA, Observatoire de Paris, Paris, France

W. B. Rossow

NASA Goddard Institute for Space Studies, New York, USA

Received 22 September 2003; revised 17 February 2004; accepted 18 March 2004; published 21 May 2004.

[1] A technique to estimate the uncertainties of the parameters of a neural network model, i.e., the synaptic weights, was described in the work of *Aires* [2004]. Using these weight uncertainty estimates, we compute the uncertainties in the network outputs (i.e., error bars and correlation structure of these errors). Such quantities are very important for evaluating any application of the neural network technique. The theory is applied to the same remote sensing problem as in the work of *Aires* [2004] concerning the retrieval of surface skin temperature, microwave surface emissivities and integrated water vapor content from a combined analysis of microwave and infrared observations over land. *INDEX TERMS:* 0933 Exploration Geophysics: Remote sensing; 3210 Mathematical Geophysics: Modeling; 3260 Mathematical Geophysics: Inverse theory; 3399 Meteorology and Atmospheric Dynamics: General or miscellaneous; *KEYWORDS:* remote sensing, uncertainty, neural networks

Citation: Aires, F., C. Prigent, and W. B. Rossow (2004), Neural network uncertainty assessment using Bayesian statistics with application to remote sensing: 2. Output errors, *J. Geophys. Res.*, *109*, D10304, doi:10.1029/2003JD004174.

1. Introduction

[2] Remote sensing requires the estimation of geophysical variables based on the inversion of indirect measurements. Neural network techniques have proved very successful in developing computationally efficient algorithms for remote sensing [e.g., *Aires et al.*, 2002b]. A rigorous scientific approach requires not only good retrieval quality, but also an estimate of the uncertainty of the retrieval (i.e., error bars plus correlation structure of the errors) [*Saltieri et al.*, 2000].

[3] One of the reasons for which such uncertainty estimates are important is that the retrieved geophysical variables are often used in a subsequent algorithm that requires an estimation of the errors and the correlation structure of these errors. For example, the variational assimilation approach [*Kalnay*, 2002; *Ide et al.*, 1997]) uses estimated uncertainties of model and observations to weight optimally both sources of information for the forecast.

[4] Reliability of the NN predictions is very important for any application. Confidence intervals (CI) have been developed for classical linear regression theory with well-established results [e.g., *Koroliouk et al.*, 1983]. For nonlinear models, such results are more recent [*Bates and Watts*, 1988], and in NN they are rarely available. Generally, only the RMS of the generalization error is provided but this single quantity is not situation-dependent. Other approaches use Bootstrap techniques to estimation such CI but they are limited by the large amount of computations that such approaches require. Recently, *Rivals and Personnaz* [2000, 2003] introduced a new method for the estimation of CI by using a linear Taylor expansion of the NN outputs (which makes traditional estimation of CI for nonlinear models a tractable problem).

[5] Our work is based on the developments of *Le Cun et al.* [1990] and *MacKay* [1992]. These studies introduced error bar estimates for neural networks using a Bayesian approach but these tools were developed and tested in artificial cases for a unique network output. In this paper, we use a slightly different approach than the more traditional "full Bayesian" method where scalar hyperparameters are estimated via the so-called "evidence" approach. A multiple output method is used in order to develop uncertainty tools for real-world applications. This method not only provides uncertainty estimates on the parameters of the neural network [see *Aires*, 2004], it can also evaluate a

variety of probabilistic quantities such as the uncertainty estimates of the network outputs.

[6] These developments are used in order to provide a new framework for the characterization and the analysis of various sources of neural network errors. In this work, we separate the errors that are due to the NN weight uncertainty and the errors from all remaining sources. We will comment on an approach to analyze in even more detail the various contributions to output errors. These errors are described in terms of covariance matrices that can be interpreted using eigen-vectors called ''error patterns'' [see *Rodgers*, 1990].

[7] These algorithmic developments are tested for the retrieval of surface skin temperature, microwave surface emissivities, and integrated water vapor content from a combined analysis of microwave and infrared observations (see *Aires* [2004] for a detailed description of this application).

[8] The theoretical computation of the predictive distribution of network outputs is developed in section 2. The developments in section 2 are used in section 3 to characterize the NN output uncertainty sources. The technique is applied to a neural network inversion algorithm for remote sensing in section 4. Conclusions and perspectives are given in section 5.

2. Predictive Distribution of Network Outputs

[9] The developments of this section, aimed at describing the distribution of the NN output (i.e., predictive distribution) and total output errors, are inspired by the Bayesian learning of neural networks chapter of *Bishop* [1996]. It is extended to the multivariate case which introduces matrix formulas instead of scalar ones.

2.1. Theoretical Derivation of the Network Output Error PDF

[10] The distribution of uncertainties of the NN output, $\boldsymbol{y}$, is given by:

$$P(\boldsymbol{y}|\boldsymbol{x},\mathcal{D}) = \int P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w}) \cdot P(\boldsymbol{w}|\mathcal{D}) d\boldsymbol{w}, \tag{1}$$

where $\mathcal{D}$ is the set of outputs $\boldsymbol{y}$ in a data set $\mathcal{B} = \{(\boldsymbol{x}^{(n)}, \boldsymbol{t}^{(n)}); n = 1,\ldots, N\}$ of N matched input/output couples. From *Aires* [2004, equations (16) and (24)], we find that this probability is equal to:

$$= \frac{1}{Z} \int e^{-\frac{1}{2}(\boldsymbol{t}-g_{\boldsymbol{w}}(\boldsymbol{x}))^T \cdot \boldsymbol{A_{in}} \cdot (\boldsymbol{t}-g_{\boldsymbol{w}}(\boldsymbol{x}))} \cdot e^{-\frac{1}{2}\triangle \boldsymbol{w}^T \cdot \boldsymbol{H} \cdot \triangle \boldsymbol{w}} d\boldsymbol{w}, \tag{2}$$

where $\boldsymbol{A_{in}}$ is the inverse of $\boldsymbol{C_{in}}$ the covariance matrix of the ''intrinsic noise'' of physical variables $\boldsymbol{y}$ and $\boldsymbol{H}$ is the Hessian matrix of the quality criterion used by the learning process (see *Aires* [2004] for more details on these two matrices). Note that all the terms not dependent on $\boldsymbol{w}$, like $-E_{\mathcal{D}}$ $(\boldsymbol{w}^{\star})$ in the work of *Aires* [2004, equation (24)], have been put together in the normalization factor Z. A first-order expansion of the neural network function $g_{\boldsymbol{w}}$ about the optimum weight $\boldsymbol{w}^{\star}$ is now used:

$$g_{\boldsymbol{w}}(\boldsymbol{x}) = g_{\boldsymbol{w}\star}(\boldsymbol{x}) + \boldsymbol{G}^T \cdot \triangle \boldsymbol{w}, \tag{3}$$

where

$$\boldsymbol{G} = \nabla|_{\{\boldsymbol{w}=\boldsymbol{w}^{\star}\}}(g_{\boldsymbol{w}}) \tag{4}$$

is a $W \times M$ matrix (M is the number of outputs). Introducing (4) into (2), and using $\varepsilon_y = (\boldsymbol{y} - g_{\boldsymbol{w}\star}(\boldsymbol{x}))$, we obtain:

$$P(\boldsymbol{t}|\boldsymbol{x},\mathcal{D}) \propto e^{-\frac{1}{2}\varepsilon_y{}^T \cdot \boldsymbol{A_{in}} \cdot \varepsilon_y} \int e^{-\varepsilon_y{}^T \cdot \boldsymbol{A_{in}} \cdot \left(\boldsymbol{G}^T \triangle \boldsymbol{w}\right)} e^{-\frac{1}{2}\triangle \boldsymbol{w}^T \cdot \left(\boldsymbol{G} \cdot \boldsymbol{A_{in}} \cdot \boldsymbol{G}^T + \boldsymbol{H}\right) \cdot \triangle \boldsymbol{w}} d\boldsymbol{w} \tag{5}$$

$$\propto e^{-\frac{1}{2}\varepsilon_y{}^T \cdot \boldsymbol{A_{in}} \cdot \varepsilon_y} \int e^{\boldsymbol{H}^T \cdot \triangle \boldsymbol{w} - \frac{1}{2}\triangle \boldsymbol{w}^T \cdot \boldsymbol{O} \cdot \triangle \boldsymbol{w}} d\boldsymbol{w} \tag{6}$$

where:

- $\boldsymbol{h} = \left[-\varepsilon_y{}^T \cdot \boldsymbol{A_{in}} \cdot \boldsymbol{G}^T\right]^T$;
- and $\boldsymbol{O} = \boldsymbol{G} \cdot \boldsymbol{A_{in}} \cdot \boldsymbol{G}^T + \boldsymbol{H}$.

The integral term in equation (6) can be simplified by:

$$(2\pi)^{\frac{dim W}{2}} |\boldsymbol{O}|^{-\frac{1}{2}} e^{\frac{1}{2}\boldsymbol{h}^T \cdot \boldsymbol{O} \cdot \boldsymbol{h}}. \tag{7}$$

We can rewrite equation (6) using this simplification to obtain:

$$P(\boldsymbol{t}|\boldsymbol{x},\mathcal{D}) \propto e^{-\frac{1}{2}\varepsilon_y{}^T \cdot \boldsymbol{A_{in}} \cdot \varepsilon_y} e^{\frac{1}{2}\varepsilon_y{}^T \cdot \boldsymbol{A_{in}} \cdot \boldsymbol{G}^T \left(\boldsymbol{G} \cdot \boldsymbol{A_{in}} \cdot \boldsymbol{G}^T + \boldsymbol{H}\right)^{-1} \boldsymbol{G} \cdot \boldsymbol{A_{in}} \cdot \varepsilon_y} \tag{8}$$

$$\propto e^{-\frac{1}{2}\varepsilon_y{}^T \cdot \left[\boldsymbol{A_{in}} - \boldsymbol{A_{in}} \cdot \boldsymbol{G}^T \left(\boldsymbol{G} \cdot \boldsymbol{A_{in}} \cdot \boldsymbol{G}^T + \boldsymbol{H}\right)^{-1} \boldsymbol{G} \cdot \boldsymbol{A_{in}}\right] \cdot \varepsilon_y}. \tag{9}$$

[11] This means that the distribution of $\boldsymbol{t}$ follows a Gaussian distribution with mean $g_{\boldsymbol{w}\star}(\boldsymbol{x})$ and covariance matrix:

$$\boldsymbol{C_0} = \left[\boldsymbol{A_{in}} - \boldsymbol{A_{in}} \cdot \boldsymbol{G}^T \left(\boldsymbol{G} \cdot \boldsymbol{A_{in}} \cdot \boldsymbol{G}^T + \boldsymbol{H}\right)^{-1} \boldsymbol{G} \cdot \boldsymbol{A_{in}}\right]^{-1}. \tag{10}$$

This covariance matrix can be simplified by multiplying numerator and denominator by:

$$\boldsymbol{G} \cdot \left(I + \boldsymbol{H}^{-1} \cdot \boldsymbol{G} \cdot \boldsymbol{A_{in}} \cdot \boldsymbol{G}^T\right) \cdot \boldsymbol{G}.$$

to obtain:

$$\boldsymbol{C_0} = \boldsymbol{C_{in}} + \boldsymbol{G}^T \cdot \boldsymbol{H}^{-1} \cdot \boldsymbol{G}. \tag{11}$$

We see that the uncertainty in the network outputs are due to (1) the intrinsic noise of the target data embodied in $\boldsymbol{C_{in}}$, and (2) the uncertainty described by the posterior distribution of the weight vector $\boldsymbol{w}$ embodied in $\boldsymbol{G}^T \cdot \boldsymbol{H}^{-1} \cdot \boldsymbol{G}$. This relation describes the fact that the uncertainties are approximately related to the inverse data density. As expected, uncertainties are larger in the less dense data space, where the learning algorithm gets less information.

2.2. Sources of Uncertainty

[12] In his paper, *Rodgers* [1990] separates the various sources of uncertainty into three components: (1) random error due to measurement noise, (2) model error due to uncertain model parameters and inverse model bias, and

(3) null space error due to the inherent finite resolution of the observing system and lack of information outside the range of the weighting functions. We think that it is difficult to characterize the sources of errors using this classification because they interact together when the inversion method uses a nonlinear model.

[13] In our equation (11), we have separated the sources of error in two terms, the ''intrinsic noise'' with covariance matrix $\boldsymbol{C_{in}}$, and the neural inversion term with covariance matrix $\boldsymbol{G^T H^{-1} G}$. Our neural inversion term refers to the errors due only to the uncertainty in the inverse model parameters and all the remaining ''outside'' sources of errors are grouped in $\boldsymbol{C_{in}}$.

[14] The inversion uncertainty can itself be decomposed into three sources, corresponding to the three main components of a neural network model:

[15] 1. The imperfections of the learning data set $\mathcal{B}$, which include simulation errors when $\mathcal{B}$ is simulated by a radiative transfer model, colocation and instrument errors when $\mathcal{B}$ is a collection of in situ and satellite colocations, null space errors, etc. This is probably the most important source of uncertainty due to the inversion technique.

[16] 2. Limitations of the network architecture because the model might not be optimum, with too few number of degrees of freedom, or a structure that is not optimal. This is usually a lower-level source of uncertainty because the network can (partly) compensate for these deficiencies.

[17] 3. A nonoptimum learning algorithm because as good as the optimization technique is, it is impossible, in practice, to be sure that the global minimum $\boldsymbol{w}^\star$ has been found instead of a local one. We think that this source of uncertainty is limited.

[18] Some of these sources of uncertainty can be assessed, for example, by performing some Monte Carlo simulations.

[19] $\boldsymbol{C_{in}}$ includes all other sources of errors. Our approach allows for the estimation of the global $\boldsymbol{C_{in}}$, but if some individual terms are known, it is possible to subtract them from $\boldsymbol{C_{in}}$. For example, if the instrument noise is known, it is possible to measure the impact of this noise on the NN outputs. The individual terms can then be subtracted from the global $\boldsymbol{C_{in}}$. For simplification and because we do not use such a priori information, we adopt the hypothesis that $\boldsymbol{C_{in}}$ is constant for each situation, only the inversion term being situation-dependent. Again, any a priori information about any nonconstant term in $\boldsymbol{C_{in}}$ could be used in this very flexible approach.

[20] Note that the specification of the sources of uncertainty by the approach of *Rodgers* [1990] uses mainly the concept of Jacobians of either the direct or the inverse model in order to linearize the impact of each error source. Linearity and Gaussian variables are easily manageable analytically, the algebra being essentially based on the covariance matrices. For example:

[21] 1. $\boldsymbol{C_M} = \boldsymbol{D_x} \cdot \boldsymbol{E} \cdot \boldsymbol{D_x^T}$, the covariance of the errors due to instrument noise, where $\boldsymbol{D_x} = \frac{\partial g_w}{\partial x}$ is the contribution function and $\boldsymbol{E} = \langle \boldsymbol{\eta}^T \cdot \boldsymbol{\eta} \rangle$ is the covariance matrix of instrument noise $\boldsymbol{\eta}$. This additional term is actually the multivariate equivalent of the expression found in the work of *Wright et al.* [2000] where the noise model is explicitly introduced in the Bayesian framework.

[22] 2. Or $\boldsymbol{F} = \boldsymbol{A_b} \cdot \boldsymbol{C_b} \cdot \boldsymbol{A_b^T}$, the covariance of the forward model errors, where $\boldsymbol{C_b}$ is the covariance matrix errors of the forward model parameter, $\boldsymbol{b}$, and $\boldsymbol{A_b}$ is the sensitivity matrix of observations $\boldsymbol{b}$ with respect to $\boldsymbol{b}$ [*Rodgers*, 1990].

[23] Some bridges can be built to link our error analysis and the approach used in variational assimilation by *Rodgers* [1990]. In the work of *Aires et al.* [2004], such Jacobians are analytically derived in the neural network framework. This makes feasible the use of Rodgers' estimates. The difference would be that our ''linearization'' uses Jacobians that are situation-dependent; this means that the estimation of the error sources would be nonlinear in nature. This will be the subject of another study.

[24] Another approach for the empirical characterization of the various sources of uncertainties is to use simulations. For example, for the instrument noise-related uncertainty, it is easy to introduce a sample of noise into the network inputs and analyze the consequent error distribution of the outputs. The advantage of such simulation approach is that it is very flexible and allows for the manipulation of non-Gaussian distributions. This will be the subject of another study.

3. Error Characterization and Analysis

[25] A neural network inversion scheme, including first guess information, has been developed to retrieve surface temperature (Ts), water vapor column amount (WV) and microwave surface emissivities at each frequency/polarization (E_m), over snow- and ice-free land from a combined analysis of microwave (SSM/I) and infrared (from International Satellite Cloud Climatology Project) data [*Aires et al.*, 2001; *Prigent et al.*, 2003a]. See *Prigent et al.* [2003b] for the snow-covered land case. The present study aims, in part, at providing uncertainty estimates for these retrievals. Both cloudy and clear-sky versions of this retrieval scheme have been developed but for simplicity, only the clear-sky case is discussed here. In this section, the technical developments of section 2 are used to characterize uncertainty sources.

3.1. Distribution of Network Outputs

[26] After the learning stage, we estimate $\boldsymbol{C_0}$, the covariance matrix of network errors $\varepsilon_y = (\boldsymbol{t} - g_{\boldsymbol{w}}(\boldsymbol{x}))$, over the database $\mathcal{B}$. Equation (11) shows that this covariance adds the errors due to neural network uncertainties and all other sources of uncertainty. Table 1 gives the numerical values of $\boldsymbol{C_0}$ for the particular example from *Prigent et al.* [2003a]. The right/top triangle is for the correlation, and the left/bottom triangle is for the covariance. The diagonal values give the variance of errors of quantity. The correlation part indicates clearly that some errors are highly correlated. This is why it would be a mistake to monitor only the error bars, even if they are easier to understand.

[27] The correlations of errors exhibit the expected behavior. Errors in Ts are negatively correlated with the other errors, with large values of correlation with the vertical polarization emissivities, for the channels that are much less sensitive to the water vapor ($E_m 19V$ and $E_m 37V$). The vertical polarization emissivities are larger than for the horizontal polarizations and are often close to one, with the consequence that the radiative transfer equation in channels that are much less sensitive to the water vapor (the 19 and

Table 1. Covariance Matrix $\boldsymbol{C_0}$ of Network Output Error Estimated Over the Database $\mathcal{B}$[a]

	Ts	*WV*	E_m19V	E_m19H	E_m22V	E_m37V	E_m37H	E_m85V	E_m85H
Ts	2.138910	−0.24	**−0.87**	**−0.72**	**−0.76**	**−0.84**	**−0.72**	**−0.49**	**−0.32**
WV	−1.392113	14.708836	0.16	−0.06	0.14	0.05	−0.15	−0.18	**−0.37**
E_m19V	−0.006294	0.003179	0.000024	**0.77**	**0.88**	**0.89**	**0.74**	**0.60**	**0.42**
E_m19H	−0.005261	−0.001143	0.000019	0.000024	**0.72**	**0.73**	**0.81**	**0.60**	**0.56**
E_m22V	−0.006274	0.003140	0.000024	0.000020	0.000031	**0.84**	**0.71**	**0.71**	**0.54**
E_m37V	−0.006121	0.001049	0.000021	0.000018	0.000023	0.000024	**0.81**	**0.70**	**0.50**
E_m37H	−0.005290	−0.002954	0.000018	0.000020	0.000020	0.000020	0.000025	**0.65**	**0.67**
E_m85V	−0.004895	−0.004945	0.000020	0.000020	0.000027	0.000023	0.000022	0.000046	**0.79**
E_m85H	−0.003906	−0.011933	0.000017	0.000022	0.000024	0.000020	0.000027	0.000044	0.000067

[a]The right/top triangle is for correlation and left/bottom triangle is for covariance; the diagonal gives the variance. Correlations with absolute value higher than 0.3 are in bold.

37 GHz channels), the radiative transfer equation is quasi-linear in *Ts* and in E_mV. In contrast, errors in water vapor are weakly correlated with the other errors: the largest correlation is with the emissivity at 85 GHz in the horizontal polarization. The 85 GHz channel is the most sensitive to water vapor and, since the emissivity for the horizontal polarization is lower than for the vertical, the horizontal polarization channel is more sensitive to water vapor. Correlations between the water vapor and the emissivities errors are positive or negative, depending on the respective contribution of the emitted and reflected energy at the surface (which is related not only to the surface emissivity but also to the atmospheric contribution at each frequency). Correlations between emissivity errors are always of the same sign and are high for the same polarizations, decreasing when the difference in frequency increases.

[28] The correlations involved in the PDF of the errors described by the covariance matrix $\boldsymbol{C_0}$ make it necessary to understand the uncertainty in a multidimentional space. This is more challenging than just determining the individual error bars, but it is also much more informative: the diagonal elements of the covariance matrix provide the variance for each output error, but the off-diagonal terms show the level of dependence among these output errors. To statistically analyze the covariance matrix $\boldsymbol{C_0}$, we decompose it into its orthogonal eigen-vectors. This base set constitutes a set of "error patterns" [*Rodgers*, 1990] so that the contribution of each of these patterns to the total error is decorrelated. In practice, the eigen-vectors in columns of $\boldsymbol{L}$ are the error patterns, $\boldsymbol{l}_k$, given by:

$$\boldsymbol{C_0} \cdot \boldsymbol{l}_k = s_k \boldsymbol{l}_k. \tag{12}$$

The eigen-vectors $\boldsymbol{l}_k$ need to be multiplied by $s_k^{\frac{1}{2}}$ so that each output component in $\boldsymbol{y}$ has the same statistical weight in the definition of the error patterns (i.e., this is the normalized PCA). The error in the network outputs is the sum of the individual errors:

$$\varepsilon_{\boldsymbol{y}} = \sum_{k=1}^{S_L} a_k \boldsymbol{l}_k, \tag{13}$$

where the factors a_k follow a Gaussian random distribution with unit variance. The interpretation of the different error patterns can provide a useful insight into the origin of the errors.

[29] Figure 1a presents the percentage of variance explained by the cumulated eigen-vectors. In this way the PCA provides an estimate of the number of degrees of freedom in the retrieval error structure. Components 1 and 2 explain 55 and 30 percent of the error, respectively. This means that the errors are concentrated in the first two "error patterns." Figure 1b shows the first four PCA components in the output variable space. The first component is essentially related to *Ts* and the emissivities in vertical polarization with a weight on water vapor close to zero: negative value for *Ts* and positive ones for the E_m, especially for the E_mV, are consistent with the correlations of errors found in Table 1 that indicate that *Ts* and E_mV errors are anticorrelated. Water vapor dominates the second PCA component, along with the emissivities for channels that are more sensitive to water vapor, namely 22 GHz and for 85 GHz horizontal polarization. Maps of the first component of the PCA for two months (not shown) do not show any well defined spatial structures that are related to surface characteristics, which is a good result. The PCA second component maps (not shown) are somewhat related to the water vapor fields, with positive value of the component in areas of large WV and negative ones in dry air regions. That suggests that the inversion tends to underestimate WV in humid regions and overestimate it in dry ones, which might be related to the use of absolute values of the humidity in the retrieval, instead of relative humidity values that would give more weight to low WV amounts. An over-representation of dry situations in learning data set can also be an explanation for the underestimation of WV in wet situations.

3.2. Covariance of Output Errors Due to the Neural Inversion

[30] We already saw in the work of *Aires* [2004] that the matrix $\boldsymbol{H}^{-1}$ is the covariance of the PDF of network weights. The use of the gradient $\boldsymbol{G}$ transforms this matrix into $\boldsymbol{G}^T\boldsymbol{H}^{-1}\boldsymbol{G}$, the covariance error of the NN outputs associated with the uncertainty of weights [*Aires*, 2004]. Note that multiplication by $\boldsymbol{G}$ regularizes $\boldsymbol{H}^{-1}$ so that for this particular purpose of the estimation of the output errors, $\boldsymbol{H}$ does not need to be regularized as described in the work of *Aires* [2004].

[31] Table 2 represents this covariance matrix $\boldsymbol{G}^T\boldsymbol{H}^{-1}\boldsymbol{G}$ averaged over the whole learning database $\mathcal{B}$. Even if some of the bottom left values representing the covariance matrix are close to zero, structure is still present in this matrix, as is shown in the correlation part (top right). This is an artifact since the variability ranges of the variables are quite different from each other. The error correlation matrix

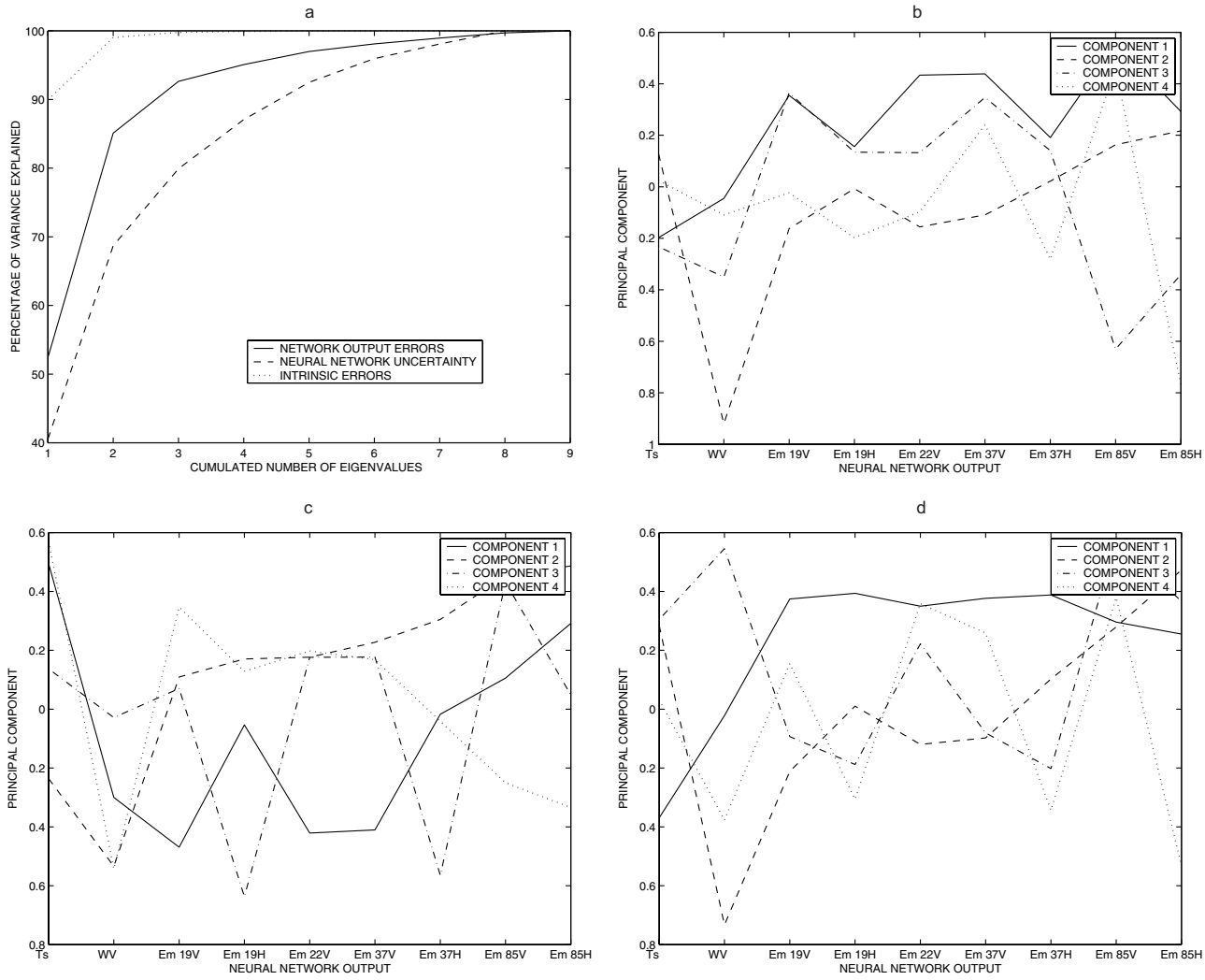

Figure 1. Eigen-decomposition of covariance matrices: (a) explained variance, (b) error patterns for $\boldsymbol{C_0}$ (network output errors), (c) error patterns for $\boldsymbol{G^T H^{-1} G}$ (errors due to neural network uncertainty), and (d) error patterns for $\boldsymbol{C_{in}}$ (intrinsic errors).

$\boldsymbol{G^T H^{-1} G}$, related to the NN inversion method, has relatively small magnitudes with a maximum of 0.55. However, it has structure similar to the global correlation matrix, with the same signs of correlation and similar relative values between the variables.

[32] As in section 3.1, we use an eigen-decomposition of $\boldsymbol{G^T H^{-1} G}$ to find the "error patterns" involved in this part of the errors. Figures 1a and 1c shows the explained cumulated variance spectrum and the corresponding error patterns. The overall behavior of the first components is rather similar to the analysis of matrix $\boldsymbol{C_0}$ of section 3.1: The first component is related to *Ts* and the emissivities in vertical polarization (negative value for *Ts* and positive ones for the E_m, especially for the E_mV). Water vapor dominates the second

Table 2. Covariance Matrix $\boldsymbol{G^T H^{-1} G}$ of Error Due to Network Uncertainty, Averaged Over the Database $\mathcal{B}$[a]

	Ts	*WV*	E_m19V	E_m19H	E_m22V	E_m37V	E_m37H	E_m85V	E_m85H
Ts	0.493615	−0.14	−0.28	−0.14	−0.25	**−0.32**	−0.16	−0.19	−0.06
WV	−0.106484	1.063071	0.10	−0.02	0.09	0.02	−0.07	−0.15	−0.25
E_m19V	−0.000325	0.000167	0.000002	**0.33**	**0.55**	**0.55**	0.28	0.27	0.08
E_m19H	−0.000255	−0.000060	0.000001	0.000006	0.26	0.22	0.29	0.10	0.13
E_m22V	−0.000268	0.000152	0.000001	0.000001	0.000002	**0.50**	0.26	0.28	0.12
E_m37V	−0.000330	0.000033	0.000001	0.000000	0.000001	0.000002	**0.34**	**0.38**	0.14
E_m37H	−0.000270	−0.000183	0.000001	0.000001	0.000000	0.000001	0.000005	0.16	0.26
E_m85V	−0.000231	−0.000282	0.000000	0.000000	0.000000	0.000000	0.000000	0.000002	**0.43**
E_m85H	−0.000128	−0.000681	0.000000	0.000000	0.000000	0.000000	0.000001	0.000001	0.000006

[a]The right/top triangle is for correlation and left/bottom triangle is for covariance; the diagonal gives the variance. Correlations with absolute value higher than 0.3 are in bold.

Table 3. Covariance Matrix $\boldsymbol{C_{in}}$ of Intrinsic Noise Errors, Estimated Over the Database $\mathcal{B}$[a]

	Ts	*WV*	E_m19V	E_m19H	E_m22V	E_m37V	E_m37H	E_m85V	E_m85H
Ts	1.645294	−0.27	**−0.99**	**−0.92**	**−0.86**	**−0.95**	**−0.88**	**−0.55**	**−0.37**
WV	−1.285629	13.645765	0.17	−0.06	0.14	0.05	−0.16	−0.19	**−0.39**
E_m19V	−0.005968	0.003011	0.000021	**0.89**	**0.91**	**0.92**	**0.83**	**0.63**	**0.46**
E_m19H	−0.005006	−0.001083	0.000017	0.000017	**0.83**	**0.86**	**0.98**	**0.71**	**0.66**
E_m22V	−0.006005	0.002988	0.000023	0.000019	0.000029	**0.87**	**0.80**	**0.75**	**0.58**
E_m37V	−0.005790	0.001015	0.000020	0.000017	0.000022	0.000022	**0.90**	**0.72**	**0.54**
E_m37H	−0.005019	−0.002770	0.000017	0.000018	0.000019	0.000019	0.000019	**0.74**	**0.76**
E_m85V	−0.004663	−0.004662	0.000019	0.000019	0.000026	0.000022	0.000021	0.000043	**0.82**
E_m85H	−0.003777	−0.011251	0.000016	0.000021	0.000024	0.000019	0.000026	0.000042	0.000060

[a]The right/top triangle is for correlation and left/bottom triangle is for covariance; the diagonal gives the variance. Correlations with absolute value higher than 0.3 are in bold.

PCA component, along with the emissivities for channels that are more sensitive to water vapor, namely 22 GHz and for 85 GHz horizontal polarization.

3.3. Covariance of the Intrinsic Noise of Target Values

[33] To estimate $\boldsymbol{C_{in}}$, we use equation (11):

$$\boldsymbol{C_{in}} = \langle \boldsymbol{C_0} \rangle_{\mathcal{B}} - \left\langle \boldsymbol{G}^T \boldsymbol{H}^{-1} \boldsymbol{G} \right\rangle_{\mathcal{B}}, \tag{14}$$

where the two right hand terms are the covariance matrix of the total output errors averaged over $\mathcal{B}$ (section 3.1) and the covariance matrix of the output errors due to the network inversion scheme averaged over $\mathcal{B}$ [*Aires*, 2004]. Table 3 gives the numerical values of the matrix $\boldsymbol{C_{in}}$: The right/top triangle is for the correlation and the left/bottom triangle is for the covariance.

[34] Intrinsic error correlations can be very large (up to 0.99). The structure of $\boldsymbol{C_{in}}$ is also very similar to the structure of the global error correlation matrix, the only noticeable difference being the larger correlation values.

[35] The eigen-decomposition shows that most of the error variability is related to the first pattern: The first component explains 90% of the errors, meaning that the number of degrees of freedom in the retrieval error variability is limited. It is mostly related to *Ts* and to the emissivities with very similar weights. As for the other matrices, maps of the PCA components do not have very particular spatial structures and are rather similar to the other maps.

3.4. Hyperparameters Optimization

[36] We saw in the work of *Aires* [2004] that the hyperparameter matrices $\boldsymbol{A_{in}}$ and $\boldsymbol{A_r}$ can be used a priori in the quality criterion for the training of the NN. This would have a regularization effect on the network. How are these hyperparameters to be obtained a priori? In a Bayesian framework, one type of estimation procedure is the so-called "evidence" approximation scheme [*Gull*, 1988; *MacKay*, 1992] based on the conventional statistics "type II maximum likelihood" [*Berger*, 1985].

[37] Another simpler approach could be to omit the hyperparameters in the first stage, by using the simplified data and regularization criteria of *Aires* [2004, equations (6) and (9)]. This is the method adopted in our study. After the learning process, the hyperparameters $\boldsymbol{A_{in}}$ and $\boldsymbol{A_r}$ can then be directly estimated and used in a new and more constrained quality criterion to re-train the NN. We can iteratively alternate the learning process and the hyperparameter estimation until the hyperparameters stabilize. It would be interesting to monitor the evolution of both of these matrices.

4. Uncertainty of Network Outputs

4.1. Network Outputs Error Estimate

[38] Once $\boldsymbol{C_{in}}$ is available, we can estimate a $\boldsymbol{C_0}(\boldsymbol{x})$ that is dependent on the observations $\boldsymbol{x}$, the term $\boldsymbol{G}^T\boldsymbol{H}^{-1}\boldsymbol{G}$ varying with input $\boldsymbol{x}$. It should be noted that the use of the regularization for matrix $\boldsymbol{H}$ presented in the work of *Aires* [2004] has virtually no consequences for the results obtained for the error bars in the following. Using no regularization for the Hessian matrix is possible since $\boldsymbol{H}$ is multiplied by the gradients in $\boldsymbol{G}^T\boldsymbol{H}^{-1}\boldsymbol{G}$. This is an additional argument that the regularization helps the matrix inversion without damaging the information in the Hessian.

[39] $\boldsymbol{C_0}(\boldsymbol{x})$ is estimated for each of the 1,239,187 samples for clear-sky pixels in July 1992. Figure 2 presents the monthly mean standard deviations (square root of the diagonal terms in $\boldsymbol{C_0}(\boldsymbol{x})$) for four outputs: the surface skin temperature *Ts*, the columns integrated water vapor *WV*, and the microwave emissivities at 19 GHz for vertical and horizontal polarizations.

[40] The errors exhibit the expected geographical patterns. Large errors on *Ts* are concentrated in regions where the emissivities are lower and/or highly variable: inundated areas and deserts. In inundated areas for instance (around the rivers like the Amazon or the Mississippi) or in coastal regions, the contribution from the surface is weaker and sensitivity to *Ts* is lower because the emissivities are lower. In sandy regions through desert areas, due to higher transmission in the very dry sandy medium, microwave radiation does not come from the very first millimeters of the surface, but from deeper below the surface, the lower the frequency the deeper [*Prigent and Rossow*, 1999]. As a consequence, the microwave radiation is not directly related to the skin surface temperature (see *Prigent and Rossow* [1999] for a detailed explanation) and *Ts* cannot be retrieved with the same accuracy. The same arguments hold for the errors in emissivity. All the parameters being tightly related for a given pixel, the water vapor errors are also rather large in inundated regions and in sandy areas.

4.2. Marginalization of the Error Probability

[41] The marginalization of the total error PDF consists in conditioning part of it by integrating over some of the error

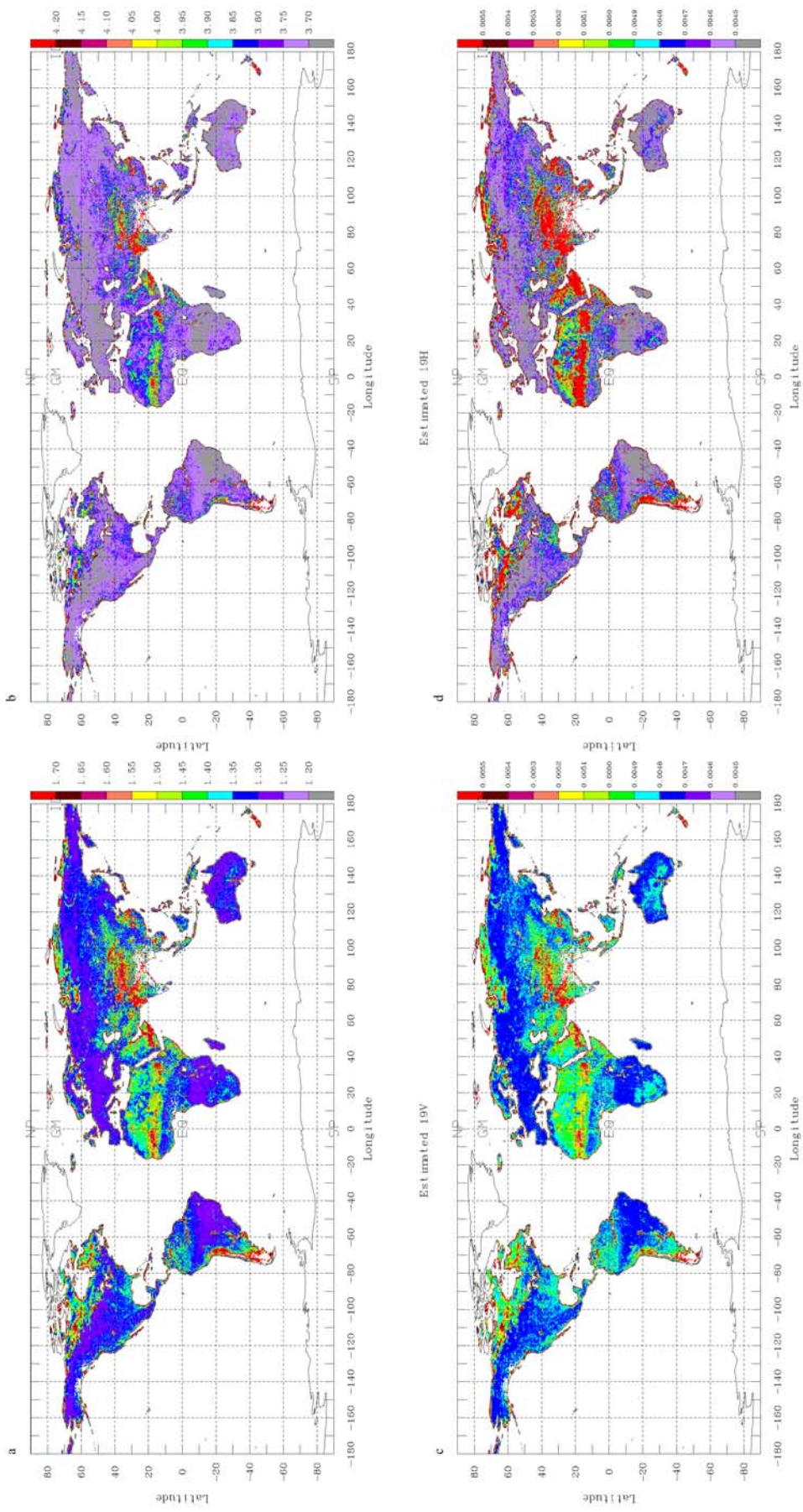

Figure 2. Standard deviation of error maps for: (a) surface skin temperature *Ts*, (b) columns integrated water vapor *WV*, (c) microwave emissivity at 19 GHz vertical polarization, and (d) microwave emissivity at 19 GHz horizontal polarization.

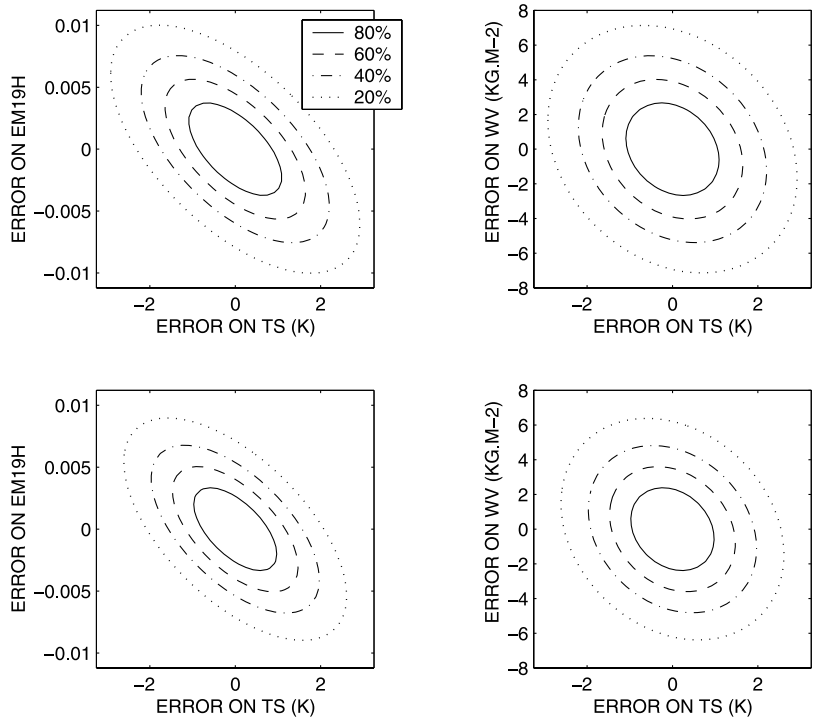

Figure 3. Two-dimensional marginalization of the network output error PDF for surface skin temperature, *Ts*, integrated water vapor, *WV*, and 19 GHz emissivity for horizontal polarization, E_m19H: Top two graphs are for deserts, and bottom two are for tropical forests. Contour lines represent the equal probability ellipsoids with levels of 80, 60, 40, and 20% of the maximum of the PDF, from the center to the outside.

variables and analyzing only the remaining few. Conditioning of the error probability is a good compromise between analyzing all the variables at the same time with the eigen-value decomposition (but each mode is only part of the variability) and observing the error bars for only one variable. In this section, the total PDF of output errors is projected onto only two variables to obtain a two-dimensional PDF of errors. This allows us to quantify the spread of errors in the same way as an histogram for a one-dimensional measurement. It also gives a measure of the correlation of errors between the two considered variables.

[42] In Figure 3, such a two-dimensional marginalization of the error PDF is presented for *Ts* paired with *WV* or E_m 19 GHz. Because the total PDF of errors is Gaussian, the contour plots of the marginalized error probability are equal probability ellipsoids. There is no bias and as a consequence, ellipsoids are centered on zero. In this figure, the data samples are separated for deserts and tropical forests. First, for both surfaces *Ts* errors are anticorrelated with emissivities, as already discussed. The probability ellipsoids for *Ts* versus *WV* are almost symmetric around 0 in *Ts* errors, meaning that errors in *Ts* and *WV* are poorly correlated. Second, the errors are more important for desert surfaces than for tropical forest, confirming the results obtained in Figure 2.

[43] It should be noted that the error estimates would be considerably improved if outliers were excluded, such as coast-contaminated or wetland pixels. However, rather than filtering difficult retrievals, we prefer to perform the retrieval for all situations as long as the error estimate is specified.

4.3. Outlier Detection

[44] What is the behavior of the neural retrieval when the situation is particularly difficult like when the first guess is far from the actual solution? In principle, the nonlinearity of the neural network allows it to have different weights on the observations and first guess information, depending on the situation. For example, if the first guesses are better in tropical cases than in polar cases, the neural network will have inferred this behavior during the learning stage, and then will give less emphasis to the first guess when a polar situation is to be inverted. This assumes once again that the training data set is correctly sampled. To understand the behavior of the uncertainty estimates better, a good strategy is to introduce artificial errors for each source of information and to analyze the resulting impact on the network outputs. The goal of this section is to validate our uncertainty estimate by analyzing extreme case, we don't investigate here physical error structures.

[45] In Figure 4, the retrieval STD error change index is presented to show the effect of perturbating the mean inputs or the mean FGs by an artificial error. The impact of these artificial errors is measured in term of percentage of the regular STD retrieval error as estimated in section 4.1. For example, an impact index of 120% means that the regular STD retrieval error estimate increases by 20% when the input is perturbed. The impact indices can be compared for

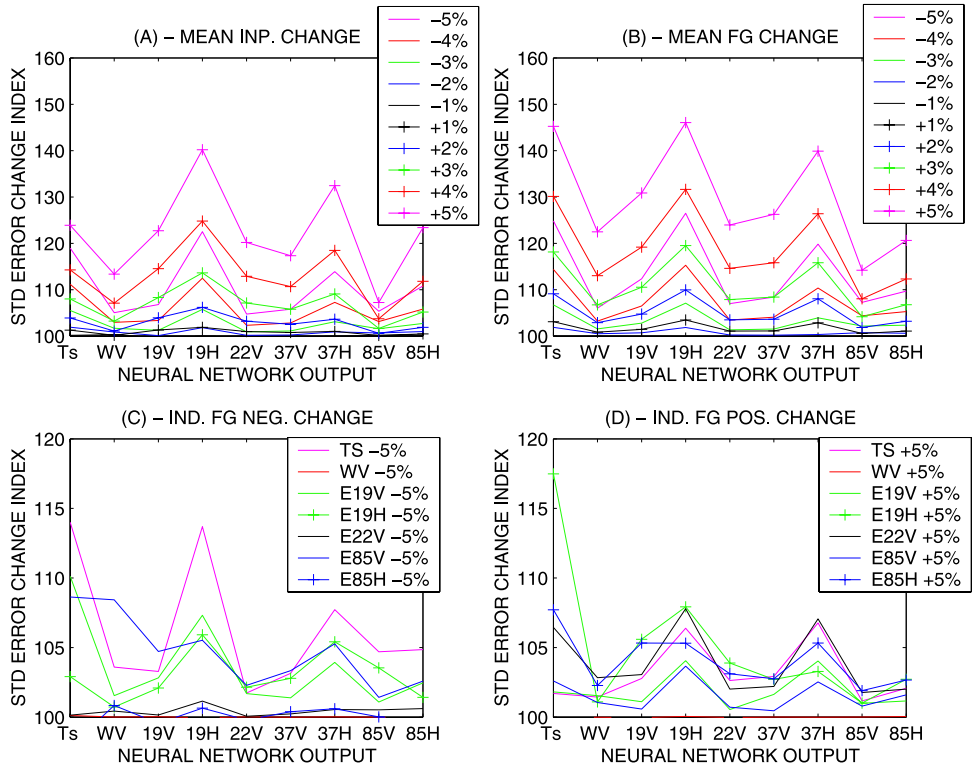

Figure 4. Estimated STD error change index for an artificial perturbation: (a) of the mean input, (b) of the mean first guess input, (c) of individual first guess negative changes, and (d) of individual first guess positive changes (see detailed explanation in the text). Statistics are performed over 20,000 samples from $\mathcal{B}$.

each of the nine network outputs. These results are obtained by averaging over the 20,000 samples in $\mathcal{B}$.

[46] Figure 4a presents the error impacts when all 17 network inputs are changed by a factor ranging from −5% to +5%. Obviously, this will introduce incoherent situations since the complex nonlinear relationships between vertical/horizontal brightness temperatures and first guesses will not be respected. As expected, the error increases monotically with the absolute value of the perturbation. However, the impact is not uniform among the output variables. For *WV*, which is retrieved with a rather low accuracy, changes in the inputs do not have a large influence. The impact on the emissivities is larger for horizontal polarizations than for vertical: horizontal polarization emissivities are much more variable than the vertical ones and as a consequence, emissivities for vertical polarization have rather similar values in outputs whatever the situation and do not depend that much on the inputs. It can also be noted that positive perturbations have a slightly stronger impact than negative ones. This is to be related to the distribution of the variables in the training data base. For the emissivities for instance, the distribution has a steep cut-off for unit emissivity, above which the emissivities are not physical. On the contrary, a large range of emissivities exists in the training data base at lower values [see *Aires et al.*, 2001, Figure 3]). As a consequence, decreasing the emissivity first guess will still be physically realistic whereas increasing it will not be.

[47] Figure 4b is the same except that the changes are made only for the first guess inputs. We note a similar behavior (nonuniform impact among output variables and with larger impact for positive perturbations) but we observe also that errors are larger than when all the inputs are perturbed in Figure 4a. This suggests that the error estimate is able to detect inconsistencies between observations and first guess inputs.

[48] In Figures 4c and 4d, the first guess input variables are perturbed individually with respectively negative and positive amplitude of 5%. For negative perturbations, the biggest impact is produced by the *Ts* first guess perturbation: it is noticeable that the *Ts* error impact is similar for the retrieval of E_m19H and for its own retrieval. For other variables, the impacts have lower levels, with almost no impact from the *WV* first guess. The *WV* first guess is associated with large error (40%) and as a consequence the NN gives little importance to this first guess. For positive individual perturbations in Figure 4d, the results are similar to the negative errors. The magnitude of the positive changes as compared to the negative ones are related again to the distribution of the variables in the training data set [see *Aires et al.*, 2001, Figure 3]: If the distribution is not symmetric around a mode value, depending on the shape of the distribution, increasing or decreasing the value can be more or less realistic.

[49] In Figure 5, "incoherencies" have been introduced between the vertical and horizontal polarizations in the brightness temperatures (*TB*) observations and in the first guess emissivities, E_ms, by increasing or decreasing one keeping the other polarization constant. In Figure 5a, we increased and decreased artificially by 5% the horizontal *TB* and in Figure 5b the same has been done for vertical

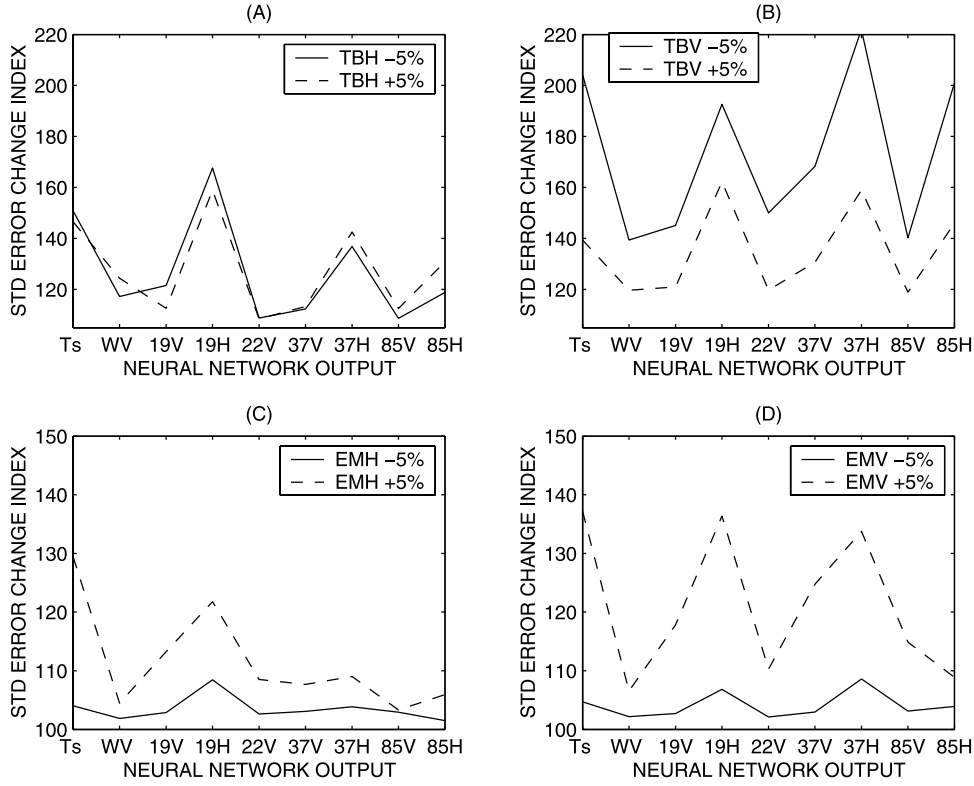

Figure 5. Estimated STD error change index for an artificial perturbation: (a) of horizontal polarization brightness temperatures, (b) of vertical polarization brightness temperatures, (c) of horizontal polarization first guess emissivities, and (d) of vertical polarization first guess emissivities. Statistics are performed over 20,000 samples from $\mathcal{B}$.

polarizations. Figures 5c and 5d are similar, for first guess emissivities instead of *TB*. Several comments can be made. First, the impact is larger for observations than for first guess errors which suggests that observations are more important for the retrieval, the first guess being used mostly as an additional constraint. Second, these polarization inconsistencies have a bigger impact than changes of the means in Figure 4. For example, the NN might emphasize the difference of polarization for the retrieval and then these inconsistencies would have a very strong impact. This shows that the NN, using complex nonlinear multivariate relationships, is sensitive to inconsistencies among the inputs. It is encouraging to see that our error estimates are able to detect such situations. Lastly, the relative impact of the positive and negative changes can be explained again by the distribution of the variables in the learning data base. For the emissivities, whatever the polarization and the frequency, the histograms are not symmetric, having a broad tail toward lower values and an abrupt end for the higher values: as a consequence, when artificially increasing the emissivities, unrealistic values are attained which is not the case when decreasing the emissivities. See *Aires et al.* [2001] for a complete description of the distributions of the learning data base and the histograms of the inputs.

[50] The results shown in Figures 4 and 5 are consistent with a coherent physical behavior, confirming that the new tools developed in this study and its companion papers can be used to diagnose difficult retrieval situations such as might be caused by bad first guesses, inconsistent measurements, situations not included in the training data set, or uncertainties of the neural network on the possible retrievals. Our a posteriori probability distributions for the neural network retrieval define confidence intervals on the retrieved quantities that allow the detection of such situations.

[51] Since outlier detection can concerns individual pertubations, in one of the measurements, another experience was done. In Figure 6, the retrieval STD error change index is presented when each of the inputs are, individually, changed to an extreme value. The FG surface skin temperature is set to 250 and 350 K. The error estimate increases, respectively, by about 25% and 90%. This unusual error estimates should allow the detection of such individual outliers. The same behavior is observed for brightness temperature measurements, or for microwave emissivity first guesses.

[52] It could be argued that a limitation of our retrieval uncertainty estimates comes from the fact that our technique is based on statistics over a data set $\mathcal{B}$. This could mean that the error estimate is only valid when we are inside the variability spanned by $\mathcal{B}$. On the contrary, it has been shown that the "local quadratic approximation" approach increases the accuracy of error estimates in sparsely sampled data space domains [see, e.g., *MacKay*, 1992].

5. Conclusion and Perspectives

[53] This paper describes a technique to estimate the uncertainties of neural network retrievals and provides a

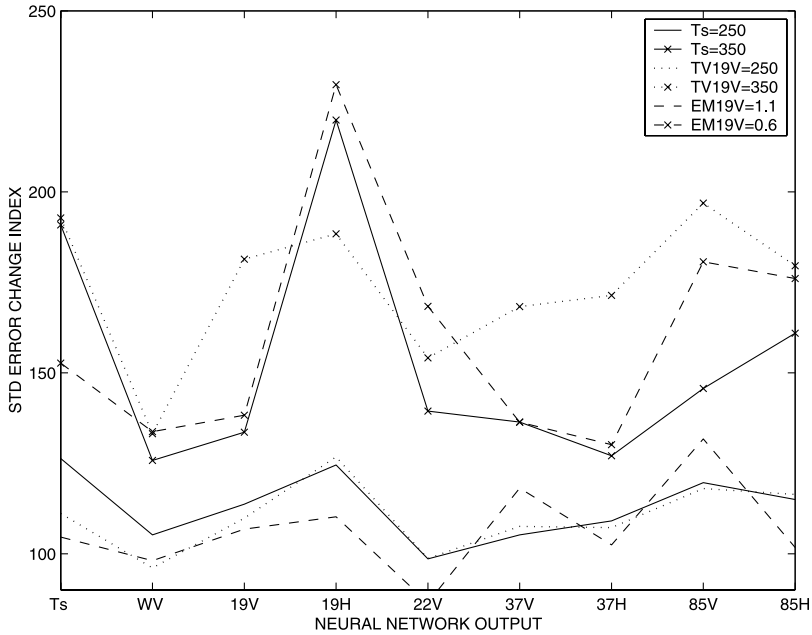

Figure 6. Estimated STD error change index for an individual perturbation of the network inputs.

rigorous description of the sources of uncertainty. The tools are very generic and can be used for different linear or nonlinear regression models. A fully multivariate formulation is introduced. Its generality will allow future developments (like the iterative re-estimation strategy or the fully Bayesian estimation of the hyperparameters). It gives insights into the neural technique that is often considered with suspicion because its mechanisms are rarely or clearly explicited.

[54] Together with the introduction of first guess information first described in the work of *Aires et al.* [2001], error specification makes the neural network approach even closer to more traditional inversion techniques like variational assimilation [*Ide et al.*, 1997] and iterative methods in general. Furthermore, quantities obtained from NN retrievals can now be combined with forecast model in a variational assimilation scheme since the error covariances matrices can be estimated. These covariance matrices are not constant, they are situation-dependent. This makes the scheme even better since it is possible now to assimilate only inversions of good quality (low uncertainty estimates). Bad situations can be discarded from the assimilation or even better can be used as an ''extreme'' detection scheme that would, for example, signal the need for an increased number of simulations in an ensemble forecast. All these new developments establish the neural network technique as a serious candidate for remote sensing in operational schemes, compared to the more classical approaches [*Twomey*, 1977].

[55] Our method provides a framework for the characterization, the analysis, and the interpretation of the various sources of uncertainty in any neural network-based retrieval scheme. This makes possible improvements in the inversion schemes. Any fault that can be detected can be corrected: Lack of data in the observation domains, errors of the model in some specific situations, or detection of extreme events. This should benefit a large community of neural network users in meteorology/climatology.

[56] Many new algorithmic developments can be pursued and we provided a few ideas. For example, the network output uncertainties can easily be used for a novelty detection (i.e., data that has not been used to train the network) or fault detection (i.e., data that are corrupted by errors, like instrument-related problems). Our determination of error characteristics can also be used with adaptative learning algorithms (i.e., learning when a small additional data set is provided after the main learning of the network has been done).

[57] We mentioned that the NN Jacobians [*Aires et al.*, 2004, 1999] can be used to express the various sources of uncertainty with even more detail, using *Rodgers*' [1990] approach. Another technical development would be the optimization of the hyperparameters as described in section 3.4 using an iterative re-estimation strategy or evidence measure in a Bayesian framework [*Near*, 1996; *Nabney*, 2002].

[58] Applications of these new tools and concepts are numerous: This approach can first be used for the inversion of satellite observations from temperature/humidity sounding instruments. The technique described in the work of *Aires et al.* [2002a] will be used to assess the quality and the difficulties in the retrieval of atmospheric profiles such as temperature, water vapor, or ozone. It would be very interesting to quantify the uncertainties for each atmospheric layer. In that sense, this will give an overview of the actual vertical resolution that can be expected with the next-generation instruments like IASI (Infrared Atmospheric Sounding Interferometer) or AIRS (Atmospheric Infrared Sounder).

[59] We would like to test how beneficial these uncertainty estimates would be when inverted satellite measurements are assimilated instead of raw brightness temperatures. Another application concerns the analysis of climate systems [*Aires and Rossow*, 2003]. In this modeling of dynamical systems, the prediction uncertainty might be used to detect complex situations where the attractor can diverge toward various basins of attraction.

Notation

- $\boldsymbol{y}$ vector of physical variables to retrieve, outputs of the NN.
- $\boldsymbol{M}$ dimension of $\boldsymbol{y}$, number of outputs in the NN.
- $\boldsymbol{t}$ target vector of physical variables in data set $\mathcal{B}$.
- $\boldsymbol{x}$ observations vector, inputs of the NN.
- η SSM/I instrumental noise, noise on inputs $\boldsymbol{x}$ of the NN.
- ε_v generic error symbol for variable v.
- P_v generic probability measure for variable v.
- $\boldsymbol{C_0}$ $(=\boldsymbol{A_0}^{-1})$, covariance matrix of total error on retrieved physical variables $\boldsymbol{y}$.
- $\boldsymbol{C_{in}}$ $(=\boldsymbol{A_{in}}^{-1})$, covariance matrix of intrinsic noise on physical variables $\boldsymbol{y}$, equivalent to $1/\beta$ in traditional Bayesian formulation.
- $\boldsymbol{C_r}$ $(=\boldsymbol{A_r}^{-1})$, covariance matrix for weight regularization, equivalent to $1/\alpha$ in traditional Bayesian formulation.
- $\boldsymbol{H}$ $= \nabla|_{\boldsymbol{w}} (\nabla|_{\boldsymbol{w}} (E_{\mathcal{D}}(\boldsymbol{w})))$, the Hessian matrix of the log-likelihood.
- $\boldsymbol{G}$ $\nabla|_{\{\boldsymbol{w}=\boldsymbol{w}^*\}} g_{\boldsymbol{w}}$.
- $\boldsymbol{C_M}$ the covariance of the errors due to instrument noise
- $\boldsymbol{E}$ $= \langle \varepsilon^T \cdot \varepsilon \rangle$, covariance matrix of the measurement errors.
- $\boldsymbol{F}$ covariance matrix of the radiative transfer model errors.
- $\boldsymbol{C_b}$ the covariance matrix of the forward model parameter errors.
- $\boldsymbol{D_x}$ $= \frac{\partial g_w}{\partial x}$ is the contribution function.
- $\boldsymbol{A_b}$ $= \frac{\partial x(y)}{\partial b}$ is the sensitivity of observations $\boldsymbol{x}$ with respect to $\boldsymbol{b}$ the parameters of the radiative transfer model.
- $\boldsymbol{L}$ matrix whose columns are the "error patterns" $\boldsymbol{l}_k$
- $.^T$ transposition operator.
- $\langle \cdot \rangle_{\mathcal{B}}$ expectation operator.
- $g_{\boldsymbol{w}}$ neural network model, or transfer function for our application.
- $\boldsymbol{w}$ $\{w_i;\ i = 1,\ldots,W\}$, the vector of the network weights.
- W dimension of $\boldsymbol{w}$.
- $\mathcal{B}$ learning database, that includes outputs $\mathcal{D}$.
- $\mathcal{D}$ target or network output database.
- N number of samples in $\mathcal{D}$ and $\mathcal{B}$.
- $E_{\mathcal{D}}(\boldsymbol{w})$ data term of the quality criterion.

[60] **Acknowledgments.** We would like to thank Ian T. Nabney for providing the Netlab toolbox from which some of the routines have been used in this work. Filipe Aires would like to thank Andrew Gelman for very interesting discussion about modern Bayesian statistics. This work was partly supported by NASA Radiation Sciences and Hydrology Programs.

References

Aires, F. (2004), Neural network uncertainty assessment using Bayesian statistics with application to remote sensing: 1. Network weights, *J. Geophys. Res.*, *109*, D10303, doi:10.1029/2003JD004173.

Aires, F., and W. B. Rossow (2003), Inferring instantaneous, multivariate and nonlinear sensitivities for the analysis of feedback processes in a dynamical system: The Lorenz model case study, *Q. J. R. Meteorol. Soc.*, *129*, 239–275.

Aires, F., M. Schmitt, N. A. Scott, and A. Chédin (1999), The weight smoothing regularisation for MLP for resolving the input contribution's errors in functional interpolations, *IEEE Trans. Neural Networks*, *10*, 1502–1510.

Aires, F., C. Prigent, W. B. Rossow, and M. Rothstein (2001), A new neural network approach including first guess for retrieval of atmospheric water vapor, cloud liquid water path, surface temperature and emissivities over land from satellite microwave observations, *J. Geophys. Res.*, *106*(D14), 14,887–14,907.

Aires, F., W. B. Rossow, N. A. Scott, and A. Chedin (2002a), Remote sensing from the infrared atmospheric sounding interferometer instrument: 2. Simultaneous retrieval of temperature, water vapor, and ozone atmospheric profiles, *J. Geophys. Res.*, *107*(D22), 4620, doi:10.1029/2001JD001591.

Aires, F., A. Chédin, N. A. Scott, and W. B. Rossow (2002b), A regularized neural network approach for retrieval of atmospheric and surface temperatures with the IASI instrument, *J. Appl. Meteorol.*, *41*, 144–159.

Aires, F., C. Prigent, and W. B. Rossow (2004), Neural network uncertainty assessment using Bayesian statistics with application to remote sensing: 3. Network Jacobians, *J. Geophys. Res.*, *109*, D10305, doi:10.1029/2003JD004175.

Bates, D. M., and D. G. Watts (1988), *Nonlinear Regression Analysis and Its Applications*, John Wiley, Hoboken, N. J.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York.

Bishop, C. (1996), *Neural Networks for Pattern Recognition*, 482 pp., Clarendon Press, Oxford, UK.

Gull, S. F. (1988), Bayesian inductive inference and maximum entropy, in *Maximum-Entropy and Bayesian Methods in Science and Engineering*, vol. 1, *Foundations*, edited by G. J. Erickson and C. R. Smith, pp. 53–74, Kluwer Acad., Norwell, Mass.

Ide, K., P. Courtier, M. Ghil, and A. C. Lorenc (1997), Unified notation for data assimilation: Operational, sequential and variational, *J. Meteorol. Soc. J.*, *75*, 181–189.

Kalnay, E. (2002), *Atmospheric Modeling*, *Data Assimilation and Predictability*, 364 pp., Cambridge Univ. Press, New York.

Koroliouk, V., N. Portenko, A. Skorokhod, and A. Tourbine (1983), *Aide-Mémoire de Théorie des Probabilités et de Statistique Mathématique*, 581 pp., Edition Mir, Moscow.

Le Cun, Y., J. S. Denker, and S. A. Solla (1990), Optimal brain damage, in *Advances in Neural Information Processing Systems*, vol. 2, edited by D. S. Touretzky, pp. 598–605, Morgan Kaufmann, Burlington, Mass.

MacKay, D. J. C. (1992), A practical Bayesian framework for back-propagation networks, *Neural Comput.*, *4*(3), 448–472.

Nabney, I. T. (2002), *Netlab: Algorithms for Pattern Recognition*, Springer-Verlag, New York.

Near, R. M. (1996), *Bayesian Learning for Neural Networks*, Springer-Verlag, New York.

Prigent, C., and W. B. Rossow (1999), Retrieval of surface and atmospheric parameters over land from SSM/I: Potential and limitations, *Q. J. R. Meteorol. Soc.*, *125*, 2379–2400.

Prigent, C., F. Aires, and W. B. Rossow (2003a), Land surface skin temperatures from a combined analysis of microwave and infrared satellite observations for an all-weather evaluation of the differences between air and skin temperatures, *J. Geophys. Res.*, *108*(D10), 4310, doi:10.1029/2002JD002301.

Prigent, C., F. Aires, and W. B. Rossow (2003b), Retrieval of surface and atmospheric geophysical variables over snow and ice from satellite microwave observations, *J. Appl. Meteorol.*, *42*, 368–380.

Rivals, I., and L. Personnaz (2000), Construction of confidence intervals for neural networks based on least squares estimation, *Neural Network*, *13*, 463–484.

Rivals, I., and L. Personnaz (2003), MLPs (mono-layer polynomials and multi-layer perceptrons) for nonlinear modeling, *J. Machine Learning Res.*, *3*, 1383–1398.

Rodgers, C. D. (1990), Characterizatioon and error analysis of profiles retrieved from remote sounding measurements, *J. Geophys. Res.*, *95*, 5587–5595.

Saltieri, A., K. Chan, and E. M. Scott (2000), *Sensitivity Analysis*, John Wiley, Hoboken, N. J.

Twomey, S. (1977), *Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurements*, Elsevier Sci., New York.

Wright, W. A., G. Ramage, D. Cornford, and I. T. Nabney (2000), Neural network modelling with input uncertainty: Theory and application, *J. VLSI Signal Process.*, *26*, 169–188.

F. Aires, Department of Applied Physics and Applied Mathematics, Columbia University/NASA Goddard Institute for Space Studies, 2880 Broadway, New York, NY 10025, USA. (faires@giss.nasa.gov)

C. Prigent, CNRS, LERMA, Observatoire de Paris, 61, av. de l'Observatoire, Paris F-75014, France. (catherine.prigent@obspm.fr)

W. B. Rossow, NASA Goddard Institute for Space Studies, 2880 Broadway, New York, NY 10025, USA. (wrossow@giss.nasa.gov)