OZONE FROM GOME DATA USING NEURAL NETWORK TECHNIQUE

Martin D. Müller[(1)], **Anton K. Kaifel**[(1)], **Mark Weber**[(2)]

[(1)] ***Center for Solar Energy and Hydrogen Research (ZSW)***

Heßbrühlstr. 21c, D-70565 Stuttgart, Germany

Email: martin.mueller@zsw-bw.de

[(2)] ***Institute of Remote Sensing (ife), University of Bremen,***

P.O. Box 330440, D-28334 Bremen, Germany

Email: mark.weber@gome5.physik.uni-bremen.de

1 INTRODUCTION

Since the launch of GOME in 1995, quite a number of physical ozone retrieval algorithms have been devised [1, 2], which rely on Differential Absorption Spectroscopy (DOAS) for determining the ozone slant column [3, 4]. These columns are then converted to total ozone values by means of an air mass factor (AMF) estimated from a combination of climatology and satellite position relative to sun and earth. Although this method has proved to be reliable and yields good results in most geographical regions, recent research has shown it to be somewhat unstable in the case of large solar zenith angles [5, 6]. We therefore present the first results of a new alternative approach to ozone retrieval, which relies on neural networks to exploit the information contained in GOME spectra.

As this method is very flexible, it can also be used for retrieving vertical ozone profiles. Current algorithms are commonly based on the work of Rodgers [7, 8], and retrieve the ozone by optimal estimation [see e. g. 9, 10]. This method is however computationally expensive, especially in the case of clouds, which have to be implemented in the forward model used. Furthermore, it requires the use and upkeep of a climatological *a priori* database with accurately known uncertainities, since these have a direct impact on the retrieved ozone profiles. A neural network, on the other hand, should be able to extract climatological and error information from its training data set, and thus does not require the explicit use of *a priori* ozone profiles.

2 DATA AND METHODS

In the frame of this work, two sets of data have so far been created: The first one consists of simulated GOMS spectra created by the GOMETRAN forward model [11] and was used for ozone profile retrivals. The second one is composed of GOME measurements collocated with total ozone columns measured from the ground. These two datasets are described in the following sections.

2.1 Simulated Data

To assess the potential of neural network ozone profile retrieval without the influence of measurement errors, it was decided to use a set of simulated radiances for initial tests. A total of 20000 spectra has been calculated for the wavelength regions 245–340 nm and 370–380 nm ($\Delta\lambda$=0,11 nm). While the former contains much ozone information from the Huggins and Hartley bands, the latter region was included as an atmospheric window for testing the influence of aerosols and ground albedo. The model calculations were based upon two global

Table 1: Selection of input parameters for the simulation of GOME top-of-atmosphere (TOA) spectra using GOMETRAN.

parameter	number/ value range	distribution type	mean	standard deviation
ozone profiles	435	white noise	–	–
total ozone [DU]	100–700	Gaussian	300	75
aerosol opt. dep.	0–1.2	exponential	0.2	–
elevation [km]	0–3	exponential	0.5	–
albedo	0–1	white noise	–	–
solar zenith angle [°]	0–90	white noise	–	–

climatologies, one provided by the Max Planck Institute (MPI) in Mainz, Germany, consisting of calculations from a 2D chemical transport model, the other one composed of sonde profiles and measurements from satellite sounders collected by Fortuin and Kelder [12]. Of these, 435 different ozone profiles with corresponding temperature and pressure profiles were selected and scaled to total ozone columns with a Gaussian distribution. Several other parameters were also varied in order to create an adequate distribution of points in the input parameter space; these are summarized in Table 1.

The resulting 20000 combinations of these parameters and atmospheric states may not all be observable in the real atmosphere, but the object of these simulations is not to provide a climatologically accurate image of the atmosphere, but rather to test the neural network's skill in data inversion. In this sense, a dataset representing more situations than actually present in any given atmosphere can be assumed to make this test even more demanding.

2.2 Total Ozone Collocations

GOME spectra used for collocating were obtained by applying the current version 2.0 of the GOME Data Processor extraction program GDP_EX [13]. The resulting ASCII stream was piped through a backend program, which Doppler-corrected the earthshine spectra and performed an interpolation of all solar and earthshine spectra to a common wavelength grid. Since the preprocessing took considerable time, all necessary information contained in the ASCII stream was then stored to a file in compact binary format for later use in collocation and ozone retrieval. To prevent irregularities in the data caused by the change of the wavelength boundary between channel 1a and 1b, which occured at 6th June 1998, data used so far were taken from the timerange 8th June 1998 to 31st May 2000.

The ground-based total ozone measurements consisted of data from Dobson and Brewer spectrometers collected and quality-controlled by the World Ozone and Ultraviolet Radiation Data Center [14, 15]. However, the data quality of the individual ground stations is still subject to considerable variations [16], therefore only data from a subset of about 100 WOUDC stations has been used for collocations [see 17]. In an effort to compensate for the sparsity of ground stations in the southern hemisphere (SH) oceans and in the polar regions, the WOUDC database has been supplemented with "virtual" stations created from gridded TOMS Version 7 data [18]. This can only be considered a preliminary solution, since recent resarch showed that ozone data from the TOMS sensor exhibits biases and higher errors in these regions as well [6, 19].

Since the WOUDC ozone values are distributed as daily mean values, the maximum temporal distance for collocations was set to 12 hours, i. e. GOME orbits from the entire day in question were collocated with a given ground ozone value. The maximum horizontal distance between the FOV center and the station location was

160 km. Only forward scan pixels were utilized, since the horizontal resolution of the result can not be noticably improved by including the backscan pixels.

The neural network test dataset was created from the collocations of six selected stations with known data quality and/or suitable location. These stations were Boulder (40°N, 105°W), Hohenpeissenberg (48°N, 11°E), Bangkok (14°N, 101°E), as well as three virtual stations located at (85°S, 135°E), (85°N, 135°E) and (80°N, 0°W). The training dataset consists of data from all other WOUDC and virtual stations.

In order to prevent over-representation of certaion total ozone values or geographical regions – especially the northern midlatitudes, where the station density is highest – in the datasets, a two-dimensional grid with bins of 10 DU times 10° latitude has been employed. The maximum number of collocations per bin was then limited to 500, whereby excess collocations were randomly selected and discarded. In this manner, the two datasets comprised 138489 training and 16977 test collocations.

2.3 Neural Network Setup

The core of our retrieval software consists of an artificial neural network simulator, a piece of software commonly referred to as "neural network". The feedforward type of neural network employed here, the *perceptron*, consists of a number of *neurons* ordered in 1-dimensional layers [20]. These layers are fully connected by weighted *synapses*. Each neuron of one layer is connected to all neurons of the adjacent layers. Ozone retrieval with such a net requires several phases:

Phase 1: Data from the training data set are normalized to the same order of magnitude and assigned to the input layer nodes. The number and type of GOME input data have been varied in the frame of sensitivity studies as described in Section 3.

Phase 2: Each data value passes a synapse as it *propagates* to the next layer, whereby it is multiplied by the weight associated with the synapse. These internal layers of the network are called *hidden layers*. Their number and size has to be determined empirically [for further discussion, see 21, 22]. Each hidden neuron sums up all incoming values and passes them through a sigmoidal transfer function, here the hyperbolic tangent. The output is again subjected to phase 2, repeating the the process until the output layer is reached.

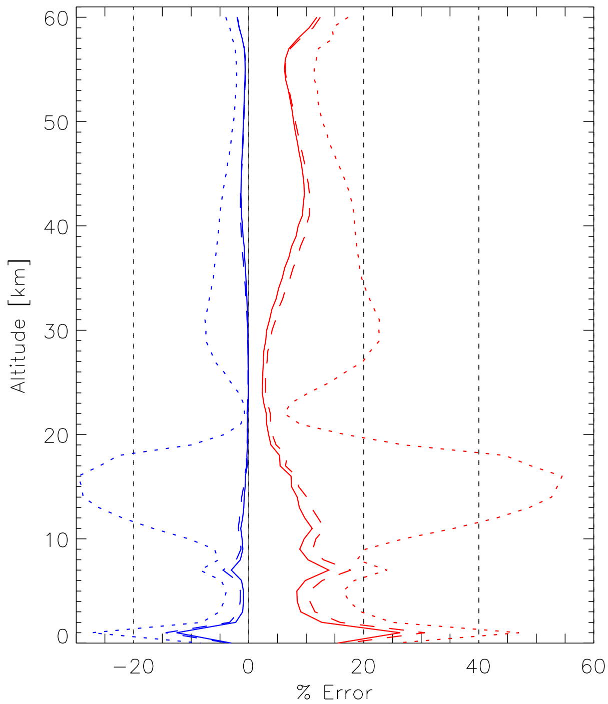
Phase 3: The output neurons determine their error by comparing their output to the collocated total ozone value or the "true" ozone profile used in the simulation of radiances. In a process called *backpropagation* this error then leads to successive modification of all weights in the network, with the aim of reducing the error in the next run. The specific *training algorithm* forcing these corrections is essential for the method's success. In our case the RPROP algorithm developed by Riedmiller and Braun [23] yielded the best results.

Phase 4: The training procedure described in phases 1-3 is repeated for all input data in the training data set, i. e. one *epoch*. To train an ozone retrieval network takes about 10^4 to 10^5 epochs. The process is considered completed once the output error with the independent test data set does no longer decrease noticably.

Phase 5: All information extracted from the input dataset by the training procedure is by now effectively stored in the neural network's weights. Due to its generalization ability the network can now be used to retrieve total ozone or ozone profiles from *all* GOME input data, not just the collocations.

3 RESULTS AND DISCUSSION

In this section preliminary results from the neural network training are being presented. Thereby, the outcome of ozone profile retrievals with simulated and total ozone retrievals with collocated data will be discussed separately.



linestyle	—	– – –	· · ·
network	1143-40-61	485-40-61	424-40-61
RMS trn.	0.096	0.116	0.448
RMS test	0.102	0.119	0.459
common input	ground albedo, SZA, elevation, wavelength range 285–330nm		
additional input	aerosol, 245–285nm, 370–380nm, p-profile, T-profile	T-profile	–

Figure 1: Bias (blue) and standard deviation (red) of neural network retrieved ozone profile with respect to the true profile used in the forward calculation. In the legend above, trn. and test refer to the neural network training and test dataset, respectively. RMS values are given in arbitrary units. Network configuration is stated in the form input-hidden-output neurons.

3.1 Simulated Ozone Profile Retrievals

The experiments carried out with the simulated data set can be roughly devided into three classes, namely the investigation of

1. sensitivity to input parameters,
2. the necessary number and size of hidden layers and
3. the achievable profile resolution

As for item 1, the most significant impact can be attributed to the temperature profile, which is partially due to ozone UV absorption and thus must have a strong correlation with the ozone profile. Since it is difficult to express the accuracy improvement for the resulting profile in a single RMS value, Fig. 1 shows the errors for three exemplary network setups. As can be seen, the temperature profile greatly decreases the errors below the ozone maximum. It was thus decided to aim at including temperature profiles – from some source still to be determined – in the retrieval process with real GOME data. Other sensitivity studies not to be discussed here in greater detail revealed that elevation, ground albedo, aerosol and solar zenith angle (SZA) did not have any noticable effect on the errors. However, it must be pointed out that some of these results are supposed to change once the transition to real data is performed – from a physical point of view, e. g. the SZA is correlated with the total AMF and thus can be expected to have some influence on the results with real data.

Training variations concerning inclusion of the pressure profile accounted for about 10% reduction in the errors, but only if the temperature profile was not used at the same time. These results suggest that the difference between solid and dashed graphs in Fig. 1 can be attributed to information contained in the additional wavelengths used.

The optimum number of hidden neurons cannot be determined exactly from the simulations. In general, the more hidden neurons a network possesses, the better it performs on the training dataset. However, at the

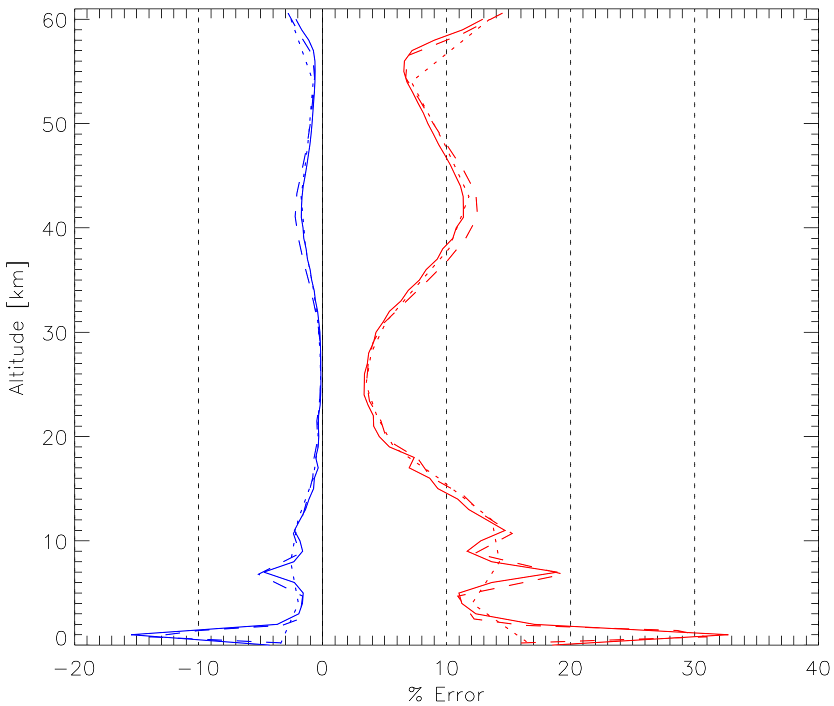


Figure 2: Comparison of Bias (blue) and standard deviation (red) for 61 (solid), 34 (dashed) and 16 (dotted) output neurons corresponding to the heights given in Table 2.

same time its generalization ability decreases due to the increased number of degrees of freedom (weights), such that the test dataset errors increase. This tradeoff forces a comprimise depended on the information content of the spectra and profiles used, and the size of the dataset available for modelling it. For the simulated data, experiments with different hidden layer configurations suggested that one layer of 20 to 40 neurons is adequate. The results did not improve when using more than one hidden layer.

As regards to the number of profile levels to be retrieved by the neural network – corresponding to the number of output neurons – the question arose if training with fewer output neurons would yield the same results as training with a greater number plus some kind of averaging procedure. Fig. 2 suggests that this is obviously not the case. From the 34 height levels present in the climatology, linear interpolation was performed to yield a 16-level and a 61-level height grid as given in Table 2. The difference between 34 and 61 levels is not significant, but any other result would have hinted at some problem of either the interpolation or the retrieval process, since linear interpolation cannot of course increase the information contained in the profiles. However, in the troposphere, averaging over the retrieved 34- or 61-level profiles would definitely lead to larger errors than obtained by the 16-level retrieval, which means a certain redundancy in present in the 34-level climatological profiles.

3.2 Total Ozone Retrievals

Total ozone was derived from GOME data by using the same wavelength range as in the operational GOME level 2 product [2], 325–335 nm, as a basis. Furthermore all neural networks were supplied with azimuth and zenith angles relative to satellite and north pole with respect to sun and the pixel center ($2^3 = 8$ angles total). Also the pixel center's latitude and the LOS viewing direction (west, nadir, east) was included in the training

Table 2: Height level boundaries used for simulated profile retrieval.

# levels	level boundaries in [km]
61	60, 59, 58, ..., 1, 0
34	60.6, 58.6, 56.5, 54.4, 52.2, 50.0, 47.7, 45.5, 43.3, 41.1, 39.0, 36.9, 34.8, 32.8, 30.9, 29.0, 27.1, 25.2, 23.4, 21.5, 19.7, 17.9, 16.1, 14.4, 12.5, 10.7, 8.8, 6.8, 4.7, 2.5, 1.9, 1.4, 0.8, 0.2
16	60, 54, 48, 43, 38, 34, 31, 28, 25, 22, 19, 16, 12, 8, 4, 0

Table 3: Results of total ozone retrievals on GOME data collocated with WOUDC measurements in the timerange 08/06/1998 to 30/04/2000 with different neural network configurations. "Best epoch" refers to the lowest test dataset error.

network	trn. epochs	trn. RMS	best epoch	test RMS	input variation
63-7-1	99999	16.30	97990	13.01	every other wavelength used
111-10-1	63075	16.27	63036	12.92	–
111-30-1	29950	15.90	29880	12.91	–
159-10-1	40161	16.54	37758	14.59	PMD data
237-12-1	34486	16.12	13422	14.49	320–340 nm, PMD data

data. Note that no cloud-clearing procedure or selection of clear pixels was employed and cloudy pixels were treated in the same way as clear ones. One hidden layer with 10 neurons was used.

Some variations from this scheme and the corresponding results are listed in Table 3. The varying number of training epochs is not to be taken too seriously: Since the error approaches its minimum asymptotically, training has been stopped in some cases as soon as changes in the error would only occur in the last digit. Overall, the standard configuration is performing well compared to the variations. A minimum configuration network with halved spectral resolution exhibits slightly larger errors, but takes less than half of the computational resources to train, whilst using 30 hidden neurons tripels the resources needed with only a small gain in the test data error. Using more input data obviously degrades the result, especially the inclusion of PMD data. Two reasons for this behaviour could be inadequate PMD error checks leading to contamination of the dataset with bad values, or the inability of the neural network to deal with the high-resolution PMD data in a useful way, such as cloud correction. These issues have to be further examined. However, since the results presented here do not in any way take account of cloud contamination, the neural network obviously derives this information from the training dataset, even without the PMDs.

4 CONCLUSION AND OUTLOOK

To summarize the results of the work presented here, it has been shown that neural networks have the potential to do real-time ozone profile retrievals from GOME data, provided a suitable collocation and/or simulation database can be set up. This is also a primary focus of ongoing work: So far, ozone profile collection and reformatting is investigated using data from the WOUDC, HALOE and SAGE. In case of too few collocations, a supplementary database consisting of GOMETRAN simulations with a realistic input atmospheric profile distribution is being calculated as a backup. Another problem with collocated profiles consists in the different height range covered by ozonesondes (0–25 km) and satellite sensors (20+ km). Approaches to counter this problem might be some sort of "partial training", or to supplement the data with climatological profiles.

Even if neural networks turn out to be less accurate than full-fledged physical retrieval methods, their real-time capability makes them valuable as a pre-scan tool for identifying interesting regions for closer examination. They could also be used to decrease the number of iterations in physical retrievals by providing more accurate *a priori* profiles.

As far as total ozone values are concerned, neural network column ozone will soon be available operationally, as is already the case with a neural network ozone product from TOVS data [17, 24].

Since the method described is very flexible, future adaption to advanced sounders and imagers like SCIAMACHY, MODIS, AIRS or IASI can probably be carried out swiftly once the profile database is set up. Also, neural networks can be trained to retrieve other atmospheric parameters or perform fast forward calculations.

Acknowledgements

The authors wish to express thanks to E. Hare for providing support with the WOUDC data, W. Thomas, K. Bramstedt and the ESA/ESRIN helpdesk for valuable tips regarding GOME preprocessing. Also, we thank the *ife* GOME team, especially J. Burrows, V. Rozanov and R. deBeek for fruitful discussions and guiding. This work is funded by the German Ministry of Education and Research (BMBF) under contract number 50EE9905 (at *ife*) and 50EE9906 (at ZSW).

References

[1] D. Diebel, J. Burrows, R. DeBeek, B. Kerridge, L. Marquard, K. Muirhead, R. Munro, and U. Platt, "Tracegas Study–Detailed Analysis of the Retrieval Algorithms Selected for the Level 1–2 Processing of GOME Data (Final Report)," European Space Agency/ESTEC, ESA Contract 10728/94/NL/CN, 1995.

[2] R. Spurr, *GOME Level 1 to 2 Algorithms Description*, Technical Note ER-TN-DLR-GO-0025, DLR, August 1996.

[3] U. Platt, D. Perner, and H. W. Paetz, "Simultaneous measurement of atmospheric CH_2O, O_3, and NO_2 by differential optical absorption," *J. Geophys. Res.*, vol. 84, 6329–6335, 1979.

[4] M. Eisinger, A. Richter, J. Burrows, and A. Piters, "Studies on DOAS ozone column retrieval from the UV and visible measurements of GOME," in *GOME Geophysical Validation Campaign Proceedings- Final Results (24–26 January, Frascati, 1996)*, ESA-WPP 108, European Space Agency, 1996, pp. 161–174.

[5] P. Stammes and R. B. A. Koelemeijer, "Error analysis of the GOME ozone column retrieval method," in *Proceedings of ESAMS '99*, Noordwijk: ESA, 1999.

[6] J.-C. Lambert and M. van Roozendael, "Investigation of pole-to-pole performances of space-borne atmospheric sensors with ground-based networks," *J. Atmos. Sciences*, vol. 56, 176–189, 1999.

[7] C. D. Rodgers, "Retrieval of atmospheric temperature and composition from remote measurements of thermal radiation," *Rev. Geophys. Space Phys.*, vol. 14 (4), 609–624, 1976.

[8] C. D. Rodgers, *Inverse Methods for Atmospheric Sounding – Theory and Practice*, Singapore: World Scientific Publishing, 2000.

[9] R. de Beek, R. Hoogen, V. Rozanov, and J. P. Burrows, "Ozone profile retrieval from GOME satellite data – Part I: Algorithm description," in *Proc. 3rd ERS Symposium*, Florence: ESA-SP 414, 1997.

[10] R. J. V. der A and H. Kelder, "Ozone profile retrieval with GOME," *Earth Observational Quarterly, ESA*, 1998.

[11] V. Rozanov, D. Diebel, R. J. D. Spurr, and J. P. Burrows, "GOMETRAN: A radiative transfer model for the satellite project GOME, The plane-parallel version," *J. Geophys. Res.*, vol. 102, 16683–16695, 1997.

[12] J. P. F. Fortuin and H. Kelder, "An ozone climatology based on ozonesonde and satellite measurements," *J. Geophys. Res.*, vol. 103, 31079–31734, 1998.

[13] S. Slijkhuis and D. Loyola, *GOME Data Processor Extraction Software User's Manual*, DLR-DFD, 1999.

[14] WOUDC, *Ozone Data for the World: Catalog of Ozone Stations and Catologue of Ozone Data for 1985 - 1989*, Tech. rep., Atmospheric Environment Service in cooperation with WMO, Downsview, Ontario, Canada, 1990.

[15] D. I. Wardle, E. W. Hare, D. V. Barton, and C. T. McElroy, "The world ozone and ultraviolet radiation data centre - Content and submission," in *Proceedings of the XVIII Quadrennial Ozone Symposium, 1996*, L'Aquila, Italy, September 1997.

[16] V. E. Fioletov, J. B. Kerr, E. W. Hare, G. J. Labow, and R. D. McPeters, "An assessment of the world ground-based total ozone network performance from the comparison with satellite data," *J. Geophys. Res.*, vol. 104 (D1), 1737–1747, 1999.

[17] A. K. Kaifel and M. D. Müller, *Bestimmung der Gesamtozonsäule mit Neuronalen Netzen und 3-D Analyse der Atmosphäre mit Dem '3I'-System*, Tech. rep., Final Report BMBF-Project 01 LO 9514/8, Center for Solar Energy and Hydrogen Research (ZSW), Stuttgart, 2000, (in German).

[18] R. D. McPeters, A. J. Krueger, P. K. Bhartia, J. R. Herman, C. G. Wellemeyer, C. J. Seftor, G. Jaross, O. Torres, L. Moy, G. Labow, W. Byerly, S. L. Taylor, T. Swissler, and R. P. Cebula, *Earth Probe Total Ozone Mapping Spectrometer (TOMS) Data Products User's Guide*, Technical Publication 1998-206895, NASA, 1998.

[19] G. E. Bodeker, J. C. Scott, and K. Kreher, "Global ozone trends in potential vorticity coordinates using TOMS and GOME intercompared against the dobson network: 1978-1998," March 2000, submitted to JGR.

[20] D. E. Rumelhart, G. Hinton, and R. Williams, "Learning internal representations by error backpropagation," in D. Rumelhart and J. McClellant, eds., *Parallel Distributed Processing, Vol. 1*, MIT Press, vol. 1, pp. 318–362, 1986.

[21] S. C. Huang and Y. F. Huang, "Bounds on number of hidden neurons in multilayer perceptrons," *IEEE Transactions on Neural Networks*, vol. 2, 47–55, 1991.

[22] S. Tamura and M. Tateishi, "Capabilities of a four-layered feedforward neural network: Four layers versus three," *IEEE Transactions on Neural Networks*, vol. 8 (2), 251–255, 1997.

[23] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The Rprop algorithm," in *Proc. of the ICNN*, San Francisco: IEEE, 1993.

[24] A. K. Kaifel and M. D. Müller, "Results of TOVS ozone retrieval with neural networks," in *Techn. Proc. 11th Int. TOVS Study Conference*, Budapest: ITWG, in press.