AUTOMATIC FEATURE SELECTION FOR COMBINED IASI/GOME-2 OZONE PROFILE RETRIEVAL

Martin D. Felder, Frank Sehnke, Anton K. Kaifel
Zentrum für Sonnenenergie- und Wasserstoff-Forschung Baden-Württemberg (ZSW), Industriestr. 6, Stuttgart, Germany

Abstract

The Neural Network Ozone Retrieval System (NNORSY) version 2 uses hierarchical neural networks to retrieve the atmospheric ozone profile from the combined spectra of the GOME-2 and IASI instruments on METOP. Formerly, we used a similar method with a single neural net to retrieve ozone from several UV/VIS sensors. The reason for originally developing a hierarchical network model lies in the overwhelming complexity of a joint retrieval paired with very high computational costs, which made it impossible for a simple 3-layer feed-forward neural net to learn the highly non-linear mapping from combined UV/VIS and IR spectra to ozone profiles.

Recent advances in computing technology have now made the hierarchical approach mostly obsolete, as our new network training software accelerates learning by a factor of more than 100 using graphics processing units (GPUs). This enables us to use Deep Neural Networks (DNNs) with five or more layers, which can learn the higher order correlations between input spectra and target profiles without a semi-automated hierarchical network construction step.

Due to these improvements in automation and speed, it was possible to implement a meta-learning approach using a state-of-the-art exploration technique named PGPE (Policy Gradients with Parameter-based Exploration; Sehnke et al. 2010), a recently developed Reinforcement Learning algorithm which has been found to have numerous advantages over evolutionary strategies. This algorithm automatically trains many networks with differing input features and architecture until the best combination is found. This leads to an improvement of retrieval accuracy especially in the troposphere and UTLS region.

In this paper we present the new approach and compare the former more physically based channel selection for GOME-2 and IASI with the results of the automated feature selection process. While many spectral channels come out as expected, there are also several interesting differences which we are going to discuss.

INTRODUCTION

The Neural Network Ozone Retrieval System (NNORSY) has been used to retrieve total ozone fields and ozone profiles from nadir sounders since several years now (Müller et al. 2002; Müller et al. 2003). It is a rather generic tool for building data-driven retrieval schemes, and has recently been applied to the joint retrieval of GOME-2 and IASI. In short, the workflow for setting up the scheme was as follows:

1) Produce collocations of GOME-2 FOVs with ozone profiles from ozone sondes and limb sounders.
2) Compute cloudiness for GOME-2 and IASI FOVs from AVHRR cloud mask. The AVHRR cloud mask provides flags for ice pixels as well, which we always treat as cloudy for our purposes. Note that the FRESCO+ cloud information (Wang et al. 2008) provided with the GOME-2 level 1b data is not used for purposes of pixel classification.
3) For each GOME-2 FOV, select the co-registered IASI FOV with lowest cloudiness.
4) Balance collocations to yield equal number of WOUDC/SHADOZ, ACE-FTS and MLS profiles by limiting the number of collocations in each 10° x 10° lon/lat bin.
5) Test data set: Generate using the same statistics, but only 10% size of the training set.
6) Create subsets by filtering all collocations using GOME-2 cloudiness ranges.
7) Train specialist neural networks on the subsets using Learn-O-Matic (Sehnke et al. 2012), then combine the results.

The filtering by cloud cover ranges is essential for exploiting as much information as possible from the troposphere. This is because the training algorithm will focus on the most prominent correlations of input features and ozone profiles first, moving to more subtle effects as training progresses. If the correspondence between spectra and tropospheric ozone is “diluted” by too many cloudy samples, the information may be drowned in noise and prevented from appearing by the regularization mechanisms in place.

DATA AND METHODS FOR NEURAL NETWORK TRAINING

Figure 1 an example low cloudiness training data set created in the manner described above. As can be seen, the collocations over icy and cloudy areas are missing, as well as over the SAA, where spectral spikes in GOME-2 band 1a cause our QA algorithms to trigger.

![Figure 1: Subset of 80162 training collocations with less than 50% clouds or ice. Red: MLS. Cyan: ACE-FTS. White: SHADOZ ozone sonde network. Yellow: Other ozone sondes gathered from the WOUDC database. Except for MLS, up to ten GOME-2 pixels may be collocated with one ozone profile in order to balance the resulting sources distribution.](image)

For each of the ~270,000 balanced collocations collected from data of the years 2009 and 2010, we assemble a feature vector from

- 4096 GOME-2 spectral channels
- 8700 IASI spectral channels
- 91 ECMWF temperature forecast levels
- ~20 scan geometry, climatological and cloud parameters.

This results in a total of 14 GB of collocation data, which is far too much for nowaday’s graphics cards’ memory, even if the specialized networks use only subsets of the data set. Hence a reduction of the number of features is necessary.
FEATURE SELECTION ALGORITHM

The novel feature selection procedure implemented in the frame of the Learn-O-Matic (Sehnke et al. 2012, and this issue) software suite uses an iterative algorithm based in adapting the weight decay parameter of the neural network training algorithm, which controls regularisation (e.g. Bishop 2006). A flowchart of the algorithm is shown in Figure 2. The underlying principle is to start with an over-regularized network (high weight decay), effectively allowing only simple models, and then reduce regularisation during gradient descent with the network training algorithm (RPROP; Riedmiller & Braun 1993) until there is no more improvement. The resulting network input layer is then pruned with a random component, to prevent discarding of diluted information. For example, if some piece of information is spread over five spectral channels, each of those will usually receive input weights with about 1/5th the magnitude a single channel containing the same information would be assigned. If a fixed threshold were used, all of these five channels might be discarded and the information lost.

![Flowchart of the algorithm](image)

Since the number of available collocations is limited, a feature selection procedure using the full resolution spectral data would still be hard pressed to extract the relevant features from all the redundant input information, massively increasing the number of feature selection iterations. Apart from this, the data would not fit into our GPU cards' memory. Hence for the purpose of the experiments reported here, only every 4th GOME-2 spectral channel and every 10th IASI spectral channel were used as input to the feature selection process.

RESULTS

The feature selection experiment was run on several data set and data source preselections. Due to the probabilistic nature of the feature selection process the results are to be treated with some care, since ideally one would draw conclusions based on many repetitions of the same experiment. Despite the massive boost of computation power through the use of GPUs, this was unfortunately not possible in the given time frame.
Figure 3 depicts the results from the standard setup using GOME-2 pixels with up to 50% cloudiness. As can be seen, many channels are selected in the Hartley and Huggins bands, because there is a lot of ozone profile information in these bands. Also, these shortwave channels tend to be rather noisy, which the algorithm tries to compensate for by selecting many more channels than would be needed to gather the handful of ozone profile degrees of freedom contained in the spectra.

Channels near band boundaries also seem to attract rather strong weights, especially the uncalibrated shortwave end of band 2b. It is possible that the network extracts information about calibration uncertainties and sensor degradation from this range, but due to the dependency on the level 1b processing algorithm it should not be used in an operational regime.

Figure 3: RMS of the neural network input weights (red bars) associated with GOME-2 spectral channels as determined by the automatic feature selection process. For clarity, an example spectrum showing the GOME-2 bands in different colors has been overlaid, and the approximate location of absorption bands is indicated with labels. Band 2a is generally not utilized by NNORSY. The ordinate scale refers only to the spectra, the RMS bars are unitless since their absolute values depend on various training algorithm and network architecture settings and cannot easily be compared between networks.

It is also reassuring that many of the features selected in GOME-2 bands 3 and 4 can be identified as belonging to H₂O and O₂ absorption regions. Especially the O₂(A)-band always commands at least one larger weight, which again hints at the importance of cloud information for the retrieval. It is less clear if some of the selections near the longwave end of GOME-2 band 2b, or possibly throughout band 2b, could be associated with aerosol extinction. For detecting such broadband extinction patterns the network would indeed have to utilize several wavelengths spread out over a larger range, but whether or not this is the case here would require a more in-depth analysis based on many repetitions of this experiment.

**Sensitivity to clouds**

In Figure 4 we compare the GOME-2 spectral weights from networks specialized on different cloudiness regimes. Figure 5 shows the IASI weights for the same networks. The most striking difference between the networks is the greatly reduced pick of GOME-2 spectral channels from bands 1a, 1b and 2b in the high cloudiness network. This is partly expected since the information content of the spectra with respect to ozone is reduced and the network may not be able to pick out tropospheric information. Interestingly, we see enhanced selection of IASI channels in this case, if not necessarily in the ozone bands. Because the IASI pixel size is much smaller than that of GOME-2, and we always co-register the least cloudy IASI pixel with a given GOME-2 pixel, the average cloudiness of IASI pixels is lower than the GOME-2 pixel cloudiness.

It is currently not quite clear why the high cloudiness network focuses on two distinct wavelengths at 237 nm and 315 nm; this may be a random effect and needs to be verified using more identical training runs. Also, when looking closely at the channels picked in the IASI water vapour continuum, it is noted that they often lie exactly on top of an absorption line. As we have reduced the spectral sampling interval for IASI by a factor of 10, it is clear that some lines will be missed because of this preselection.
Figure 4: As Figure 3, but comparing weights of a specialized network for low cloudiness (upper panels) with weights from a high cloudiness network. In both cases the bar heights have been scaled relative to the maximum of all spectral weights for the given network.

Figure 5: As Figure 4, but for the IASI sensor.

Figure 6 shows the selection of cloud features for three networks trained on different cloudiness ranges. Interestingly, the AVHRR ice fraction within both GOME-2 and IASI pixels seem to be of great interest in all cases. This is probably because our “cloudiness” includes ice cover, so the network has to learn a significant correction function depending on the actual ice amount detected. Cloud cover itself is not selected, possibly because this information is readily available from the spectra.
themselves. This shows that including computed quantities into the feature vector may be beneficial in some cases, but ignored in others where there are alternative, more direct sources of information. As a tendency, more cloud features are used in the high cloudiness regime. Why the GOME-2 cloud fitting mode is prominent in low cloudiness regimes but less so in high cloudiness is currently unclear. Probably it is serving as a proxy for some information upon which the FRESCO+ algorithm decides which fitting mode to use. This however is highly algorithm version dependent and will not be used in an operational setting.

Figure 6: Selection of cloud parameters for different cloudiness ranges. The horizontal bars indicate RMS weights for the given input feature. Features whose name starts with “GOME-2” stem from the FRESCO+ algorithm delivered with the GOME-2 level 1b data. The other features are derived from the AVHRR cloud mask.

Combined retrieval data sources sensitivity

The discussion so far was based on experiments using both GOME-2 and IASI spectra in the input. For comparison, we plot the GOME-2 channel 1 and 2 weights for a GOME-2 only retrieval in Figure 7, with the corresponding panel from Figure 3 repeated for ease of comparison.

Figure 7: Comparison of weights for GOME-2 channels 1 and 2 between experiment setups with and without IASI spectra entering the neural network. Channels 3 and 4 are not shown due to insignificant differences.

Without IASI data, the network uses more information from band 1a, and 1b, but even more so from the atmospheric window and aerosol channels in band 2b. However, the overall retrieval RMSE on the test data set increases considerably when depriving the network of IASI information. Table 1 elaborates on this fact, while adding comparative figures for other input data combinations. All of these experiments used the same automatic feature selection algorithm. It should be noted that the temperature profile has a big influence here because it has a high density of information regarding the ozone profile, as compared to the sensor spectra where the information is more spread out and redundant. Note also that while the GOME-2 only retrieval (#2) yields better results than IASI only (#3), the combination with temperature profiles works better for IASI (#6) than for GOME-2 (#5), since the temperature profile is more strongly correlated with ozone in the stratosphere, where GOME-2 also yields the most information. As can be gathered from the last table column, IASI reduces the RMSE considerably in the UTLS region when combined with the other sources.
<table>
<thead>
<tr>
<th>#</th>
<th>ECMWF-T</th>
<th>IASI</th>
<th>GOME-2</th>
<th>relative RMSE</th>
<th>relative RMSE UTLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>x</td>
<td></td>
<td>136%</td>
<td>181%</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>x</td>
<td></td>
<td>154%</td>
<td>225%</td>
</tr>
<tr>
<td>4</td>
<td>x</td>
<td>x</td>
<td></td>
<td>126%</td>
<td>159%</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>x</td>
<td></td>
<td>119%</td>
<td>147%</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>x</td>
<td></td>
<td>111%</td>
<td>113%</td>
</tr>
</tbody>
</table>

Table 1: Relative change of retrieval RMSE with respect to test data set collocations, compared to the standard configuration using all available data sources (row 1). The first RMSE column refers to the average over all output height levels (0-61 km), the UTLS column refers to the maximum error detected between 7 km and 12 km, usually where the steepest gradients are seen outside the tropics. ECMWF-T denotes temperature profiles from the ECMWF forecast system as used in the operational GOME-2 level 2 processor at EUMETSAT.

CONCLUSIONS

In this paper we have demonstrated that our automatic feature selection procedure is able to pick out physical information from spectral input data. While the interpretation of single input features is sometimes difficult due to the probabilistic nature of the feature selection algorithm and the highly redundant input spectra, it seems clear that no other interpretation could explain the findings. Still, we are currently running multiple experiments to improve the robustness of these results.

Furthermore, the retrieval results are comparable to the manual feature selection used previously, although as of now a direct comparison is not feasible due to the accumulated changes in the preprocessing and training software. Automatically constructed Deep Neural Nets output by the Learn-O-Matic tool replace manually assembled hierarchical nets used previously (Felder et al, in preparation), and thus facilitate the construction of new retrieval schemes tremendously.

For the next generation NNORSY operational processor we plan to combine the automated selection process with a reasonable manual pre-selection of physically relevant features to gain the best of both worlds. Once the final configuration is trained and the data reprocessed, an independent validation by BIRA is foreseen.

As is the case with all data mining applications, the quality of results improves with the size of the training data set. There is currently also the limit imposed by GPU memory size, but as graphics cards manufacturers move into the high performance computing market we expect this situation to be ameliorated soon.

The application of our feature selection and learning framework is of course not restricted to retrievals, but can be applied to any problem domain where a suitable pattern data set can be assembled.
REFERENCES


