



Physics-Informed Deep Learning for OH(3–1) Airglow Spectra Temperature Retrieval

Harry Ballington¹ and Christoph Kalicinsky¹

¹Institute for Atmospheric and Environmental Research, University of Wuppertal, Germany

Correspondence: Harry Ballington (ballington@uni-wuppertal.de)

Abstract. The GRIPS-N instrument maintains one of the longest-running records of mesospheric temperatures, derived from OH(3-1) emission spectra in the near infrared. This study presents two deep neural network architectures for automating the classification and temperature prediction of these spectra, trained on 11 years (2014–2024) of data manually classified by an expert. A dual-branch one-dimensional convolutional classifier separates degraded spectra from those suitable for temperature derivation, achieving a precision of 99.2% on a test set of 35 456 spectra. A physics-informed temperature model, in which a dense encoder extracts three effective peak intensities that are converted to temperature via the Boltzmann equation, achieves a mean absolute error of 4.17 K and a median error of 2.80 K. Both models use approximately 50 000 trainable parameters and train in minutes on a consumer GPU. The work serves as an informative resource and reference for the application of deep learning in signal processing across a range of remote sensing applications in atmospheric science.

1 Introduction

Ground-based OH* airglow observations and the derivation of temperatures from these measurements constitute a valuable tool for monitoring the upper mesosphere-lower thermosphere (MLT) region from multiple perspectives. First, they enable long-term trend evaluation (e.g. Beig, 2011a; French et al., 2020, and references therein), which is increasingly important in the context of a changing climate that produces divergent temperature responses across different altitude regimes. Second, these observations provide insight into solar influences on the Earth's atmosphere within the MLT (e.g. Beig, 2011b; French et al., 2020). Third, they permit the detection of atmospheric waves and other periodic temperature fluctuations spanning a wide spectrum of periods, from small-scale gravity waves and tides to planetary waves with periods ranging from days to weeks (see Wüst et al., 2023, and references therein). These waves are crucial for transporting energy from the surface to the upper atmosphere and constitute the primary drivers of large-scale global circulation (Holton, 1983; Andrews et al., 1987; Fritts and Alexander, 2003). In addition to capturing and monitoring these processes, airglow observations offer advantages over satellite measurements for long-term monitoring because ground-based instruments can be maintained throughout their operational lifetimes. Such observations are conducted at numerous sites worldwide, providing a globally distributed network (e.g. Bittner et al., 2002; Mulligan and Lowe, 2008; French and Mulligan, 2010; Schmidt et al., 2013; Holmen et al., 2014; Perminov et al., 2014; García-Comas et al., 2017; Kim et al., 2017).

The time series of OH* (3,1) rotational temperatures recorded from Wuppertal represents one of the longest continuous records,



with measurements initiated in the early 1980s, continuous observations beginning in mid-1987, and ongoing operation today (e.g. Bittner et al., 2002; Kalicinsky et al., 2016, 2024). These data have been obtained with two successive Ground-based Infrared P-branch Spectrometer (GRIPS) instruments: the original GRIPS-II (Bittner et al., 2002) and its successor GRIPS-N, implemented after detector failure and subsequent reconstruction (Kalicinsky et al., 2016, 2024). To ensure data continuity, the temperature-derivation methodology has remained unchanged since its inception and is implemented in a semi-automated workflow that incorporates several selection rules for spectra. For example, only spectra with a signal-to-noise ratio exceeding six are retained, and those contaminated by clouds, street-level mercury lamps, or other interferences are discarded (Bittner et al., 2002). Nightly mean temperatures are computed only when at least 25 spectra (corresponding to approximately 1.5 hours of observation) are available. Prior to GRIPS-N, the requirement was one hour for high-quality spectra or two hours otherwise (Bittner et al., 2002). Although the semi-automated pipeline now performs most processing steps, visual inspection of spectra remains an integral component of the workflow as it was before (Bittner et al., 2002). Recent developments have opened new possibilities for further enhancing the analysis. Following the advancement of modern techniques in deep learning, and by leveraging over a decade of classified spectra, we are able to develop a more sustainable and efficient method for characterising and filtering the spectra, as well as for rapid derivation of the corresponding temperatures. In this work, an algorithm based on deep neural networks is constructed, allowing for automatic classification and filtering of OH spectra, as well as a custom-tailored network that leverages the Boltzmann relations to derive the temperature from the three main peaks in the measured spectra. This work touches on both the merit and importance of the specific data used here, and lays out an example framework for how modern techniques in deep learning can be applied to modernise signal processing in atmospheric research.

The paper is structured as follows: Sect. 2 describes the GRIPS instrument and the analytical temperature retrieval from a historical perspective. Sect. 3 presents the data preprocessing, dataset construction, and the classification and temperature regression model architectures. Results from the classification and temperature prediction are presented and discussed in Sect. 4, and conclusions are given in Sect. 5.

2 Instrument and Data

2.1 The GRIPS Instrument

The GRIPS-N instrument is a Czerny–Turner spectrometer equipped with a thermoelectrically cooled InGaAs detector (Kalicinsky et al., 2016). Excited hydroxyl (OH^*) molecules in the upper mesosphere/mesopause region, concentrated in a layer at approximately 87 km altitude with a thickness of roughly 9 km (full width at half maximum) (Baker and Stair Jr., 1988), emit radiation in the near infrared. The instrument measures the emissions of the P1(2), P1(3), and P1(4) lines of the OH(3-1) band between 1.524 μm and 1.543 μm , from which rotational temperatures representative of the emission layer can be derived (e.g. Bittner et al., 2000, 2002). The instrument has collected data from Wuppertal (51° N, 7° E) since 2014 after the instrument reconstruction and the final set-up of GRIPS-N (Kalicinsky et al., 2016). Fig. 1 shows the complete times series of the GRIPS-N nightly means from 2014 until the end of 2024.



As with all real measurement data, sources of uncertainty require consideration. Data gaps arise primarily due to cloud cover, which is the dominant cause of missing observations (Kalicinsky et al., 2016). Contamination from artificial light sources, including a mercury (Hg) line at 1529.6 nm that contaminates the P1(3) emission, must also be accounted for and the corresponding spectra have to be discarded (Bittner et al., 2002; Schmidt et al., 2013). Gravity waves in the mesosphere introduce temperature fluctuations on the timescale of several minutes (Offermann et al., 2011; Kalicinsky et al., 2016), possibly adding short-period variability to the measured signal which can be interpreted as atmospheric noise.

2.2 Temperature Retrieval

The following provides a general overview of the analytical treatment for deriving rotational temperatures from OH emission spectra, as described in more detail by Sivjee and Hamwey (1987) and Bittner et al. (2000). The derivation is based on the Boltzmann distribution, which governs the population of rotational energy levels in the OH molecule. When an excited OH* molecule relaxes from a rotational state with quantum number J' in an upper vibrational level v' , it emits radiation at a characteristic wavelength. The intensity I of a given emission line is related to the rotational temperature T_{rot} by

$$I = N_{v'} A \frac{2(2J' + 1)}{Q_{v'}} \exp\left(\frac{-F}{k_B T_{\text{rot}}}\right), \quad (1)$$

where A is the Einstein coefficient of spontaneous emission, F the rotational term value of level J' , $N_{v'}$ the population of the upper vibrational level, $Q_{v'}$ the state sum, and k_B the Boltzmann constant (Meinel, 1950; Schmidt et al., 2013). The degeneracy factor $(2J' + 1)$ reflects the number of magnetic substates available to each rotational level. Since the population $N_{v'}$ and partition function $Q_{v'}$ are not directly accessible from a single measurement, three emission lines are observed and their intensity ratios can be used to eliminate these unknowns (Schmidt et al., 2013). In practice the formula is rearranged by defining $g_j = (2J'_j + 1) \cdot A_j$ for each of the three emission lines $j \in \{1, 2, 3\}$ corresponding to P1(2), P1(3), and P1(4), and taking the logarithm of Eq. (1):

$$\ln\left(\frac{I_j}{g_j}\right) = -\frac{F_j}{k_B T_{\text{rot}}} + \text{const}, \quad (2)$$

so that T_{rot} can be determined from the slope of a linear fit through the three points $(F_j, \ln(I_j/g_j))$ in the Boltzmann plane (Mies, 1974; Sivjee and Hamwey, 1987; Bittner et al., 2000). The known Einstein coefficients and term values for each line are taken from Mies (1974) and Krassovsky et al. (1962), with values $A_{\text{P1}(2)} = 16.742 \text{ s}^{-1}$, $F_{\text{P1}(2)} = -42.0 \text{ cm}^{-1}$; $A_{\text{P1}(3)} = 20.367 \text{ s}^{-1}$, $F_{\text{P1}(3)} = 32.9 \text{ cm}^{-1}$; $A_{\text{P1}(4)} = 21.823 \text{ s}^{-1}$, $F_{\text{P1}(4)} = 138.4 \text{ cm}^{-1}$. Since the OH(4-2) R-branch overlaps with the OH(3-1) P-branch an iterative correction is additionally applied (Lange, 1982; Schmidt et al., 2013).

Fig. 1 shows the complete data set of GRIPS-N nightly mean temperatures in the time interval 2014 to 2024. As noted in the introduction, the mesopause temperature is highly variable, with fluctuations on timescales from minutes to years. Visual inspection reveals a clear annual cycle which is the major part of the seasonal variation together with a semi-annual and a ter-annual cycle (Bittner et al., 2000). Additionally, long-term variations such as an approximately 11-year modulation that coincides with the 11-year cycle of solar activity, also visible in the number of sunspots and the solar radio flux F10.7 cm, and potential long-period fluctuations are present (Kalicinsky et al., 2016, 2024).

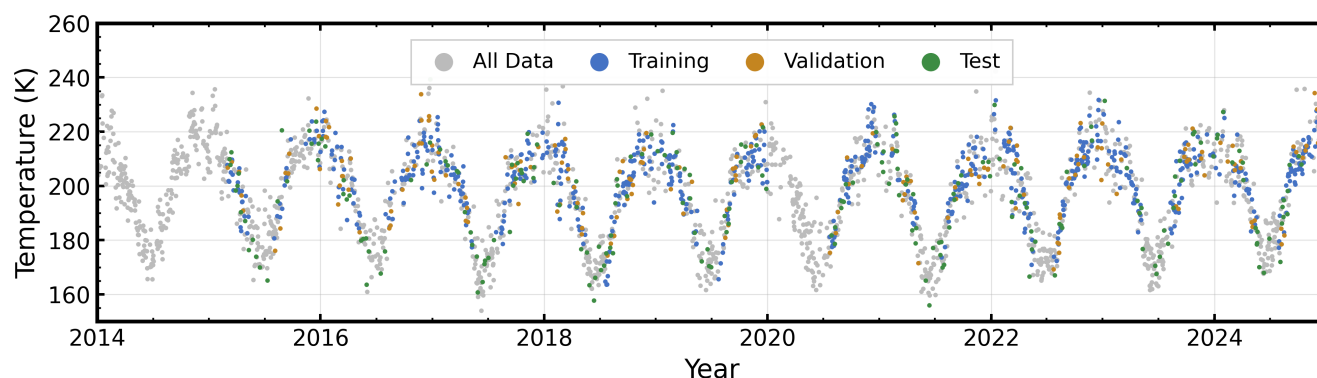


Figure 1. Long-term time series of nightly mean OH* rotational temperatures derived from GRIPS measurements at Wuppertal.

90 3 Methods

3.1 Data Preprocessing

The raw dataset is comprised of 11 years of GRIPS-N spectral observations (2014–2024), where each measurement is an emission spectrum of 301 intensity values across the OH(3-1) wavelength range. Before any model training, several filters are applied to the dataset. First, rare cases with spectra containing negative intensity values are removed, which are caused by instrument readout errors. Second, spectra contaminated by a mercury light source in the lab, which is turned on in rare cases, are identified and discarded. The lamp produces a broad emission feature near 1529 nm that overlaps with the P1(3) line. By comparing the mean intensity in the mercury emission window to the intensity baseline away from the OH peaks, with a threshold ratio of 2.0, these cases can be safely pruned. Third, specific time periods during which the instrument showed slightly increased noise characteristics, as for example in early 2020, are excluded.

100 3.2 Dataset Construction

The filtered dataset is split into training, validation, and test sets on a nightly basis, with nominal 70/20/10% split fractions. The nights are randomly shuffled to avoid bias toward any particular time interval. All spectra from a given night are assigned to the same set to avoid information leaking between splits. The training and validation sets are further restricted to nights containing at least 100 spectra classified as good, thus, nights with rather good measurement conditions. The test set retains all nights for a fair comparison against real conditions. This shifts the effective splits for train, validation, and test to ~64/16/20%. Fig. 1 shows all data colour coded for their individual use. Training on labelled data with a strong contrast between favourable and unfavourable conditions (mostly due to cloud contamination) offers two advantages. First, it sharpens the model's ability to discriminate between the two classes. Second, it places intermediate cases along a more continuous gradient between extremes, yielding output probabilities that better reflect the prediction confidence.



110 After filtering, the training, validation, and test sets comprise 132 334 spectra across 733 nights, 34 430 spectra across 191 nights, and 35 456 spectra across 226 nights, respectively. In the training and validation sets, the class balance is approximately 83/17 good/bad, whereas the test set is approximately 60/40 due to the inclusion of nights with fewer good spectra. Example good and bad-labelled spectra are shown in Fig. 2. To account for the class imbalance during classifier training, inverse-frequency class weighting is applied so that both classes contribute equally to the loss function. It should be noted that the

115 figure shows only a single case of a poor quality spectrum, and there exist a wide range of nuanced cases where a spectrum may not suitable for temperature retrieval.

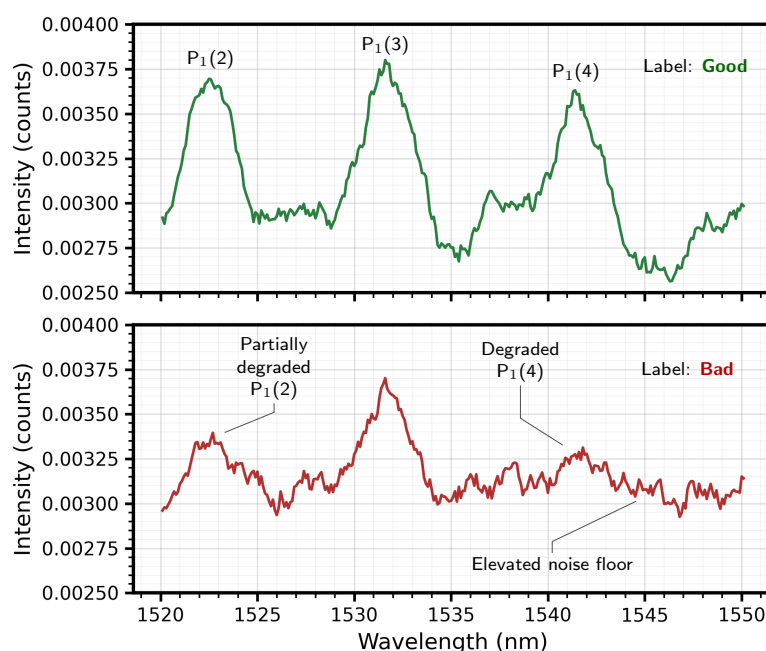


Figure 2. Example good and bad labelled intensity spectra. The good-labelled spectrum shows clear, resolvable OH(3-1) peaks, whereas the bad-labelled spectrum shows peak degradation and increased noise.

3.3 Classification Model

The classification task is to determine whether a given spectrum is of sufficient quality for downstream temperature derivation. Each input spectrum is first min-max normalised to the range $[0, 1]$, removing absolute intensity variations that depend on atmospheric conditions and instrument gain rather than spectral quality. Since the classification decision depends on local

120 shape features in the spectrum, such as the height and width of individual peaks, the presence of artefacts, and the overall noise level, a one-dimensional convolutional neural network (1D CNN) is employed. Convolutional layers apply a learned kernel that slides across the spectrum, recognising patterns at each position without needing to re-learn the same feature at every wavelength independently.



125 The architecture uses a dual-branch design. The fine branch uses a small kernel size of 5 (covering approximately 1 nm), making it sensitive to sharp features such as peak degradation, noise spikes, and narrow artefacts. The broad branch uses a kernel size of 21 (approximately 4 nm), which captures the peak envelope and inter-peak structure in a manner analogous to the Savitzky–Golay smoothing used in the analytical processing pipeline. Each branch is comprised of two convolutional blocks, where each block applies a 32-filter convolution, followed by batch normalisation, a ReLU activation, and max pooling. Batch
130 normalisation rescales the activations between layers, which encourages a stable distribution during training. Max pooling halves the sequence length at each stage, progressively broadening the effective receptive field of subsequent layers. The outputs of both branches are concatenated and passed through a final convolutional block with 64 filters. A global average pooling layer then condenses the spatial information into a single summary vector of fixed length, making the model insensitive to the exact position of spectral features. Dropout with a rate of 0.3 is applied before the final sigmoid output to prevent the
135 network from becoming overly reliant on any particular set of learned weights.

The model is trained with binary cross-entropy loss, the Adam optimiser with a learning rate of 10^{-3} , a batch size of 128, and early stopping on the validation loss with a patience of 10 epochs. The learning rate is reduced by a factor of 0.5 if the validation loss does not improve for 5 consecutive epochs. The hyperparameters were determined empirically, and training completes in several minutes on a consumer GPU. The classifier has approximately 50 000 trainable parameters. The model
140 architecture is shown in Fig. 3(a).

3.4 Temperature Regression Model

As described in Sect. 2.2, the rotational temperature is determined by the relative intensities of the three P1 lines through the Boltzmann relation (Eq. 2). Traditionally, these intensities are extracted through a series of signal processing steps: background subtraction, peak fitting, and a recursive correction for the P1(4)/P2(4) overlap (Schmidt et al., 2013). Here, the signal process-
145 ing chain is replaced with a neural network encoder that learns to extract three effective peak intensities directly from the raw spectrum, which are then passed through a fixed layer implementing the Boltzmann relation (Eq. 2) to yield temperature.

The encoder architecture is fully connected, consisting of a normalisation layer, dense layers of 128, 64, and 32 units with ReLU activations, and a final layer of 3 units with softplus activations. The softplus activation ensures strictly positive output values, which is required since they serve as inputs to the logarithm in Eq. (2). The output values do not correspond to true
150 physical peak intensities but serve as intermediate values which, when passed through the Boltzmann formula, produce a real predicted temperature. The distinction is relevant because the P1(4) line overlaps with the P2(4) line, and the traditional analytical pipeline requires a recursive correction to separate the two peaks. Since the neural network learns only effective peak intensities, this effect is avoided entirely.

The Boltzmann layer implements Eq. (2) as a fixed, non-trainable operation. It computes $y_j = \ln(I_j/g_j)$ for the three effective peak intensities, then performs a least-squares linear fit through the points (F_j, y_j) , and finally returns $T_{\text{rot}} = -1/(k_B \cdot m)$ where m is the slope. Since the physics of converting peak ratios to temperature is well understood, directly mapping the learned effective peak intensities is both simpler and more efficient.

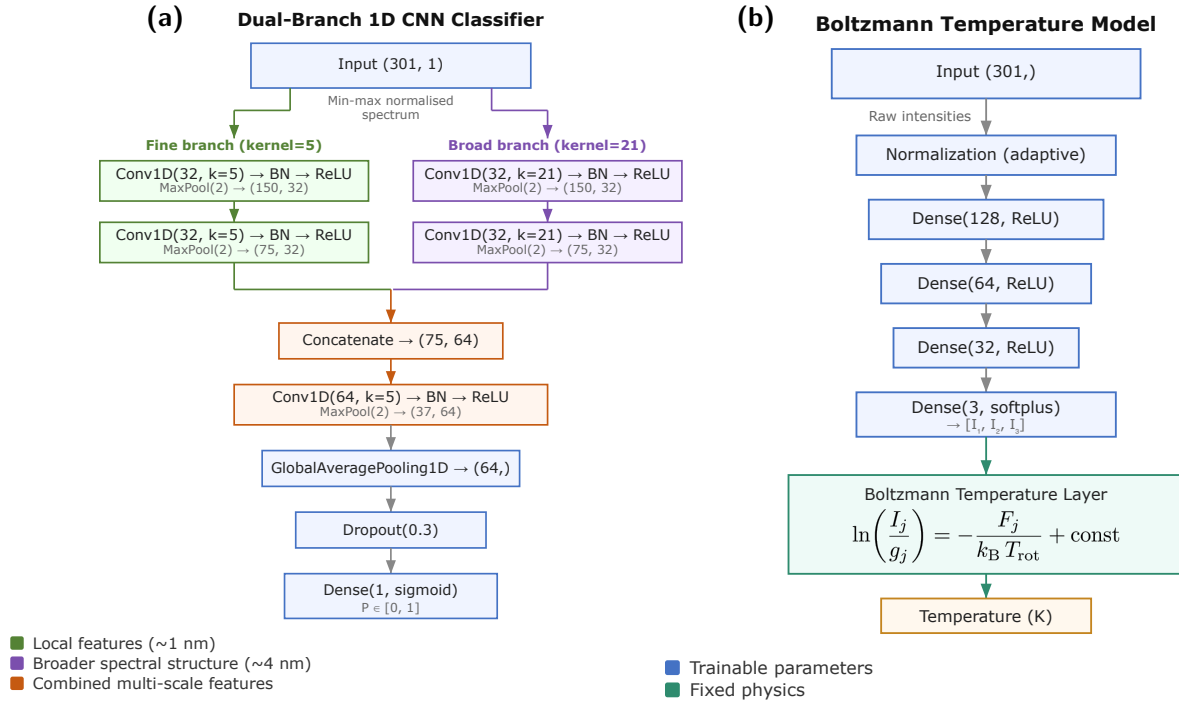


Figure 3. Architecture diagrams for the two neural network models. Parenthesised tuples denote tensor dimensions at each stage. Layers are written as LayerType(units, activation), where k is the convolution kernel size and BN denotes batch normalisation. (a) Dual-branch 1D CNN classifier. The input is a spectrum with shape (301,1), where the second dimension is the single input channel required by 1D convolution layers. (b) Boltzmann temperature model. The input is a spectrum with shape (301,), passed through fully connected (Dense) layers to extract three effective peak intensities, which are converted to temperature via the Boltzmann equation.

Unlike the classifier, the temperature model uses unnormalised intensity values as input, since the temperature depends on the relative magnitudes of the raw peak intensities. A normalisation layer adapted to the training data statistics is used as the first layer of the encoder to bring the input values into a suitable numerical range for training. For added robustness with respect to outliers, the model is trained with Huber loss ($\delta = 10\text{K}$), which behaves quadratically for small errors and linearly for large errors. The Adam optimiser is used with a learning rate of 10^{-3} , a batch size of 128, and early stopping on the validation mean absolute error with a patience of 15 epochs. The number of trainable parameters in the model is ~ 50000 . The model architecture is shown in Fig. 3(b).



165 4 Results

4.1 Classification Performance

The classifier is evaluated on the unseen test set of 35 456 spectra from 226 nights, which includes all nights regardless of data quality. For this study, *precision* is the key metric of interest, which represents the fraction of model-predicted good spectra which are truly good. False positives are the most important cases to avoid because they contaminate the downstream temperature retrieval with low quality data. The classification threshold is therefore set at $P \geq 0.9$, which is chosen to minimise false positives and consequently maximise precision, whilst retaining a good quantity of accepted spectra per night for computing a reliable nightly mean temperature.

Fig. 4(a) shows the confusion matrix at this threshold. The model achieves a precision of 99.2%, corresponding to 135 false positives out of 17 509 accepted spectra. Of the truly good spectra in the test set, 81.4% are retained after the $P \geq 0.9$ cut (recall).

Fig. 4(b) shows the distribution of predicted probabilities for good and bad spectra, where a probability of 0 indicates a confident prediction of bad and 1 indicates a confident prediction of good. The distribution is strongly bimodal, with the majority of predictions concentrated near the extremes, indicating that the model has learned to distinguish good from bad spectra with high confidence in most cases. The peak near zero is slightly sharper than the peak near one, suggesting that strongly degraded spectra are easier to classify with certainty than high-quality spectra. Intermediate cases where the model is less certain about whether the spectrum is good or bad typically correspond to spectra with mild degradation of the OH peaks, where the signal is present but partially obscured by noise from cloud cover or other sources such as street lamps in the presence of haze. These borderline cases indicate that the model could be improved with further architectural refinements.

4.2 Temperature Prediction Performance

The temperature model is evaluated on 19 639 good spectra from the test set with associated temperatures in the physically valid range. Fig. 5(a) shows the spread of predicted versus actual temperature. Although a large fraction of the data is concentrated around 200 K to 220 K, the model is able to predict extreme temperatures with very little visible systematic bias.

Fig. 5(b) shows the distribution of signed prediction errors. The model has a mean bias of 0.16 K, standard deviation of 6.22 K, and the 90th percentile error is 9.40 K. These uncertainties are well below the maximum of expected uncertainties in the case of atmospheric noise and atmospheric fluctuations. The standard deviation for the temperatures for GRIPS-II and also GRIPS-N during one night typically lies in the range from about 15 K during quiet times intervals up to more than 30 K in times with large wave activity (Offermann et al., 2011). The standard deviation is the sum of atmospheric noise, instrument noise, and atmospheric fluctuations such as waves (Offermann et al., 2011) and can be seen as an upper limit for the noise uncertainty in quite time intervals. For comparison, an unconstrained dense baseline model ($\sim 681\,000$ parameters, no analytical Boltzmann layer) was trained on the same data. The Boltzmann model achieves slightly superior mean performance with less than 10% parameters. This implies a significantly lighter training cost, and the model showed more stable behaviour in the tails of the error distribution due to the physics-informed architecture.

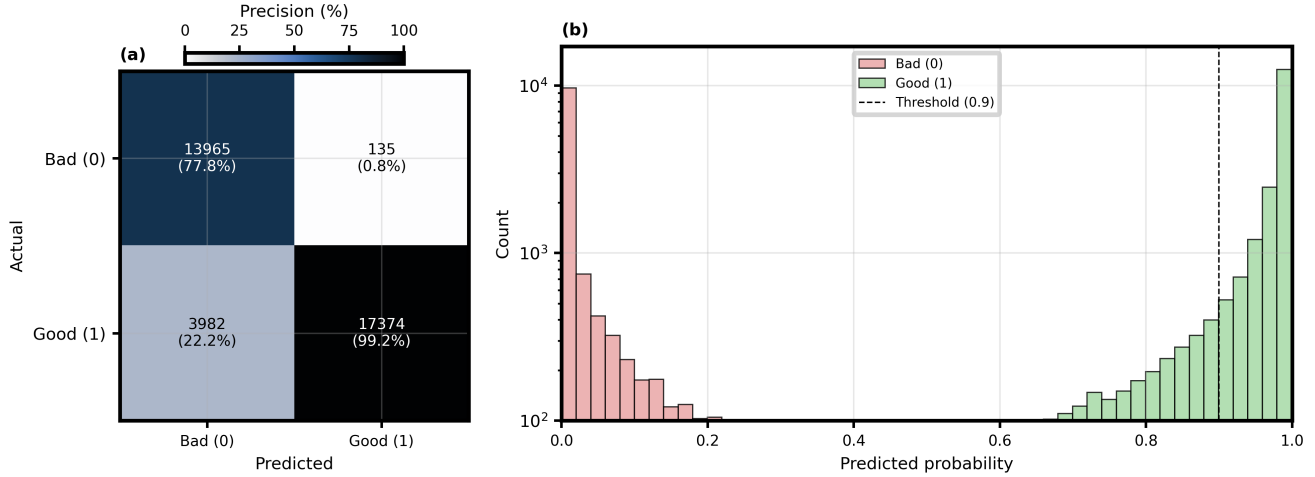


Figure 4. Classifier performance on the test set (35 456 spectra, 226 nights) at the operational threshold $P \geq 0.9$. (a) Confusion matrix, with percentages normalised per predicted class so that the diagonal cells represent the precision of each predicted class. (b) Distribution of predicted probabilities for good (green) and bad (red) spectra on a logarithmic scale. The dashed line indicates the operational classification threshold.

The temperatures reported here are computed on a per-spectrum basis. Nightly mean temperatures are obtained as the arithmetic mean of the individual predictions. An alternative approach, in which all spectra from a given night are first averaged and the temperature is then derived from the resulting mean spectrum, was also tested. However, the per-spectrum approach was found to produce better agreement with the labelled dataset.

Fig. 6 shows the effective peak intensity ratios extracted from the model's intermediate layer. As indicated in the sixth network layer in Fig. 3(b), the input to the Boltzmann layer consists of three values $[I_1, I_2, I_3]$, whose ordering is fixed such that I_1 corresponds to P1(2), I_2 to P1(3), and I_3 to P1(4). The ratios I_2/I_1 and I_3/I_1 can therefore be plotted as abstract effective values learned by the model, acting as proxies for the true peak intensity ratios. The ratio space shows a smooth temperature gradient, indicating that the network has learned a physically meaningful internal representation consistent with the theory of Sect. 2.2.

To evaluate the accuracy of the model in predicting relative temperature changes, which is an important quantity when studying wave activity in the atmosphere, the predicted temperature deltas are compared to the actual temperature deltas on both a per-spectrum and per-night basis. Fig. 7(a) shows the per-spectrum concordance scatter, where consecutive spectra within each night are compared, and Fig. 7(b) shows the per-night concordance, where nightly mean temperatures are compared between consecutive nights. The directional prediction of the model is accurate to 91.5% on a per-spectrum basis and 97.3% on a per-night basis. Table 1 presents the cumulative per-spectrum directional accuracy for selected magnitude thresholds. For temperature changes of 2 K or greater, the model predicts the correct direction 93.8% of the time. This rises to 96.0% for changes of 5 K or greater, 97.9% for 10 K or greater, and 99.1% for 20 K or greater.

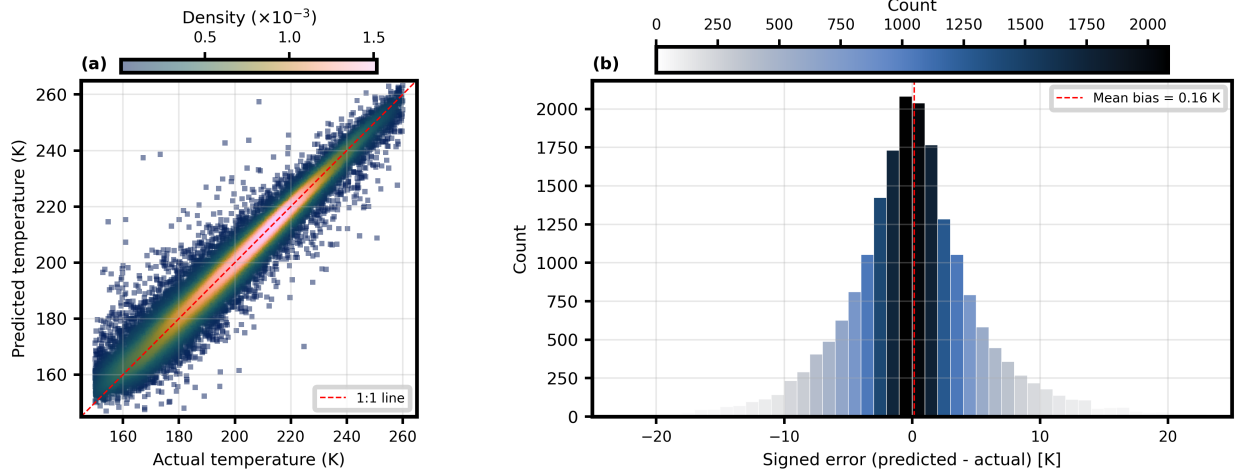


Figure 5. Temperature model performance on the test set (19 639 good spectra). (a) Predicted versus actual temperature, coloured by kernel density estimate. MAE = 4.17 K, median error = 2.80 K. (b) Distribution of signed prediction errors, coloured by count. The mean bias is 0.16 K with a standard deviation of 6.22 K.

Table 1. Cumulative per-spectrum directional accuracy by magnitude of actual temperature change. For a given threshold, the accuracy indicates the fraction of consecutive spectrum pairs where the model correctly predicts the direction of change for all pairs with $|\Delta T_{\text{actual}}| \geq$ the stated value.

$ \Delta T_{\text{actual}} \geq$	Accuracy (%)	n
2 K	93.8	18 169
5 K	96.0	16 433
10 K	97.9	13 579
20 K	99.1	8 708

As a final test on data not seen by the network during training, Fig. 8 shows the nightly mean temperature for the entire 2025 calendar year, comparing the network prediction with the legacy pipeline. The expected seasonal trend is clearly visible, with a maximum during the winter months and a minimum in the summer. The nightly mean absolute error is 0.18 K and the nightly bias is -0.01 K. This uncertainty is far below the statistical uncertainty of a typical nightly mean temperature for the legacy pipeline which is in the range of 1 to 3 K depending on the number of single measurements that enter the nightly mean (Bittner et al., 2000, 2002; Oberheide et al., 2006).

The reference data are computed from nights with at least 20 spectra classified as good according to the criteria described in Sect. 3.2. The two methods do not always retrieve a temperature from the same set of spectra: in some cases the first or last spectra of the night are taken from the network but always excluded in the legacy pipeline as they may contain fluctuations or noise because of the starting and ending process of the measurements, and in others the legacy pipeline rejects spectra because

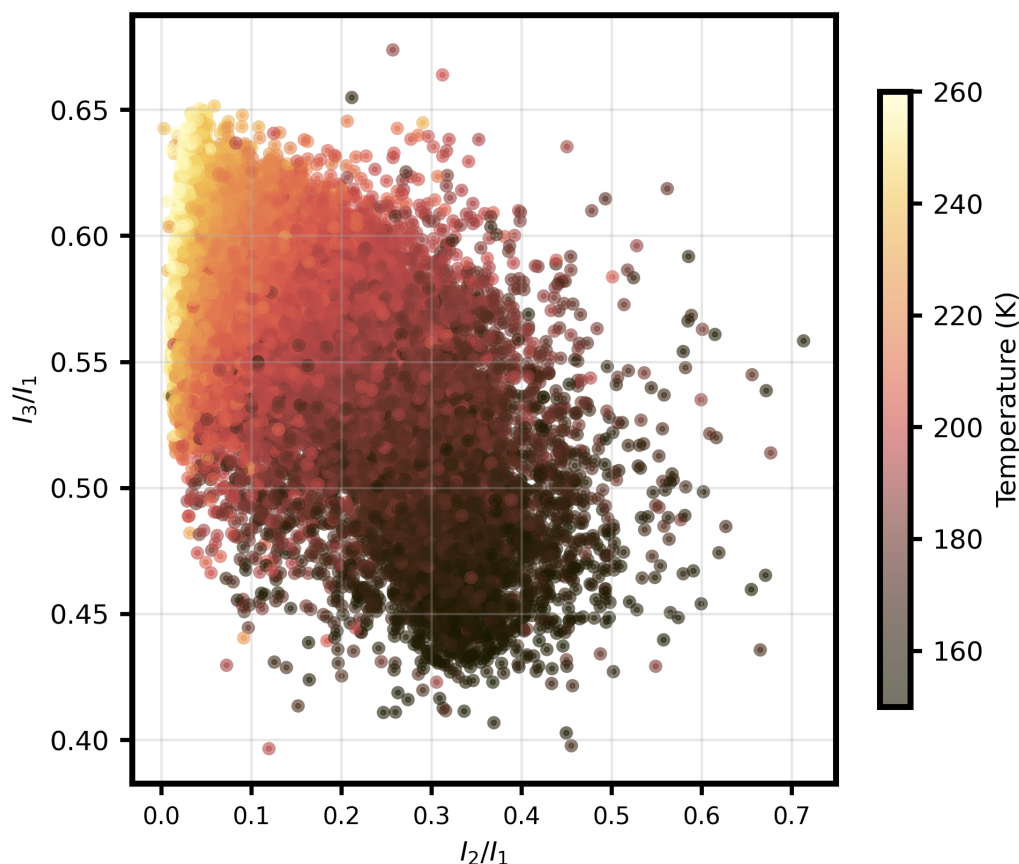


Figure 6. Effective peak intensity ratios I_2/I_1 and I_3/I_1 extracted from the model's intermediate layer, colour-coded by actual temperature. The smooth temperature gradient confirms a physically consistent learned representation.

the recursive correction for the P1(4) overlap does not converge. For a fair comparison, the nightly means shown in Fig. 8 are computed only from spectra which both methods classified as valid. It is visually clear, especially from the residual plot, that the network prediction can be used as a substitute for the legacy pipeline, providing a modern and efficient method for the retrieval of mesopause temperatures from OH airglow spectra.

230 The results shown in both Fig. 7 and Fig. 8 demonstrate that the uncertainty in the retrieved temperature relative to the legacy pipeline is sufficiently small to support both per-spectrum directional analysis and per-night temperature analysis. The framework used has general applicability to a wider range of spectral analysis techniques in atmospheric science.

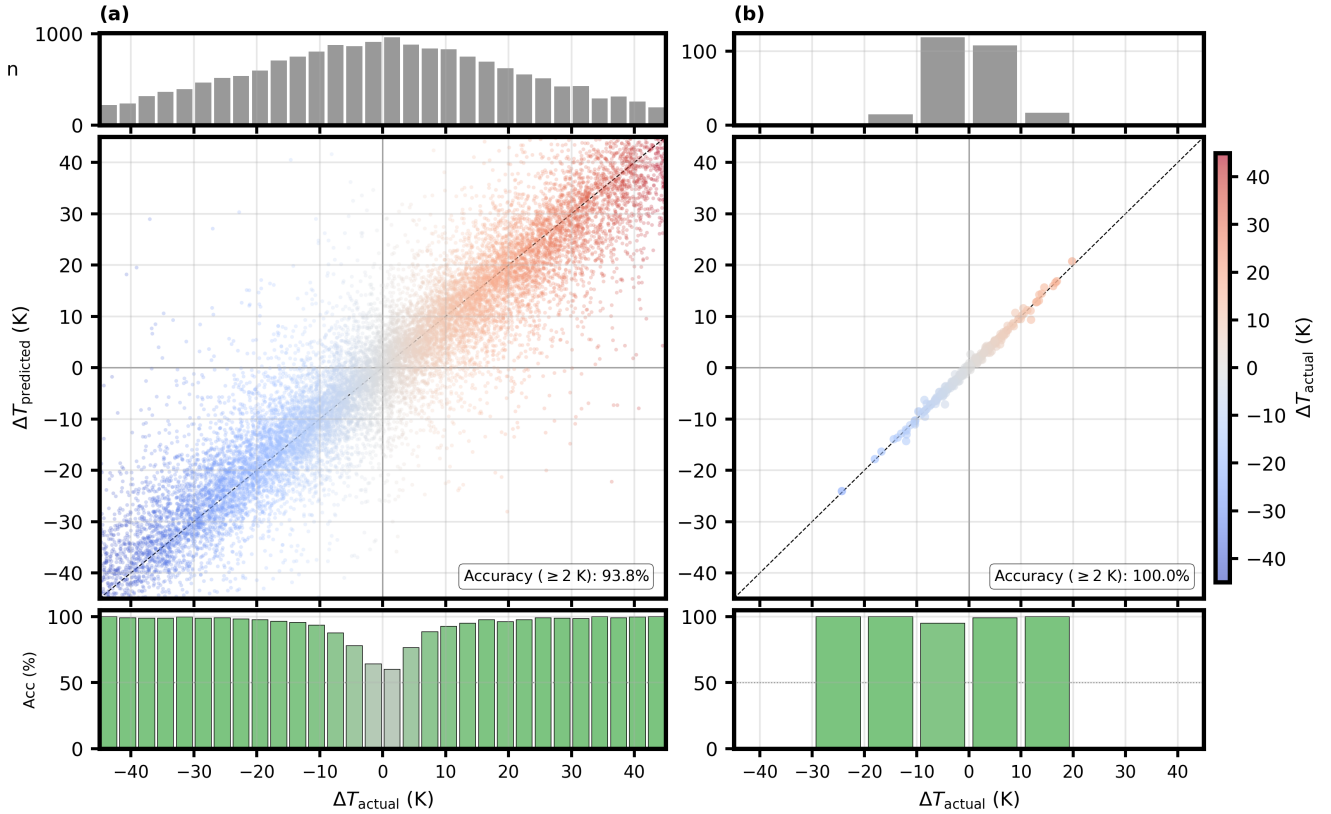


Figure 7. Directional accuracy of the temperature model. (a) Per-spectrum: predicted vs actual temperature change between consecutive spectra within each night. (b) Per-night: predicted vs actual change in nightly mean temperature. Upper panels show the actual temperature change count distribution, lower panels show directional accuracy of the network per bin.

5 Conclusions

The GRIPS-N instrument underpins one of the longest-running experimental datasets for measuring temperature trends in the mesopause, a highly variable region of the atmosphere and an indicator of anthropogenic effects on the upper atmosphere. Recent advancements in deep learning make it possible to automate and modernise the traditional signal processing pipeline for classification and temperature prediction of OH emission spectra.

Using 11 years of GRIPS-N data (2014–2024), comprising over 202 220 spectra of 301 wavelength points each (Sect. 3.2), this study presents two neural network architectures for spectral classification and temperature prediction. The dataset is split at the night level into training, validation, and test sets, with nights randomly distributed across the full observation period to avoid temporal bias.

The classifier (Sect. 3.3) uses a dual-branch one-dimensional convolutional architecture operating on min-max normalised spectra. A narrow, 1 nm-wide kernel detects small scale spectral features, while a broader, 4 nm-wide kernel captures the

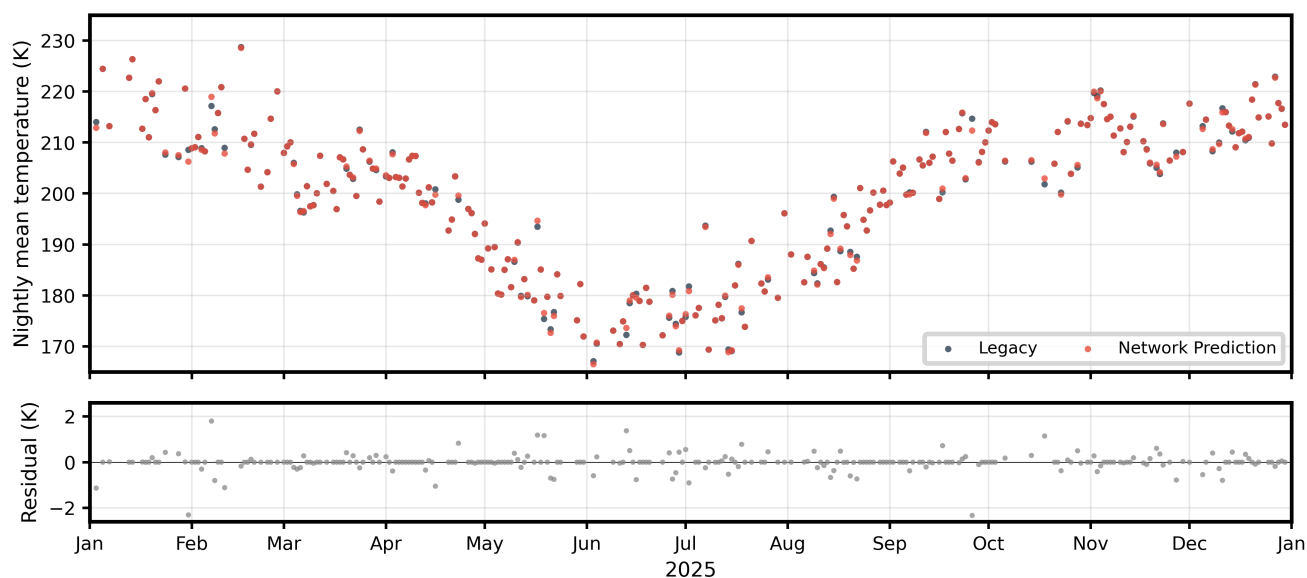


Figure 8. Nightly mean temperature for the 2025 calendar year, derived from the legacy pipeline (dark) and from the network prediction (red). The residual between the two retrievals is shown in the lower subfigure.

smoother spectral structure. On a test set of 35 456 spectra from 226 nights, the classifier achieves a precision of 99.2% at an
 245 operational threshold of $P \geq 0.9$ (Sect. 4.1), ensuring that false positives are minimised so that bad spectra do not contaminate
 the downstream temperature retrieval.

The temperature model (Sect. 3.4) takes unnormalised spectra as input and uses a dense encoder to extract three effective
 peak intensities, which are passed through a fixed Boltzmann layer that implements the known analytical relationship between
 peak intensity ratios and rotational temperature (Eq. 2). On the test set of 19 639 good spectra, the model achieves a mean
 250 absolute error of 4.17 K, a median error of 2.80 K, and a 90th percentile error of 9.40 K. The directional accuracy of the model
 in predicting whether the temperature increases or decreases is 91.5% on a per-spectrum basis and 97.3% on a per-night basis
 (Sect. 4.2).

Both models use $\sim 50\,000$ trainable parameters and train in minutes on a consumer GPU. This work illustrates the potential
 for deep learning techniques in atmospheric signal processing and facilitates the continued operation of one of the longest-
 255 running experimental records of mesospheric temperature trends.

Author contributions. H. Ballington designed the network architectures and implemented all code (Methodology, Software), and carried out
 the training, analysis, and visualization (Investigation, Formal analysis, Visualization, Writing – original draft preparation). C. Kalicinsky
 conceived the study (Conceptualization), curated the historical GRIPS-N dataset and legacy pipeline (Data curation), contributed instrument



and retrieval-methodology expertise, and validated the model results and training-data filtering strategy (Methodology, Validation). Both
260 authors reviewed and edited the manuscript (Writing – review and editing).

Competing interests. The authors declare that they have no conflict of interest.

Code and data availability. The neural network training and inference code is available at <https://github.com/hballington12/grips-n-nn>. The underlying GRIPS-N spectra are available from the corresponding author upon reasonable request.

Acknowledgements. The research by CK was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) —
265 519284835.



References

- Andrews, D., Holton, J. R., and Leovy, C. B.: Middle Atmosphere Dynamics, Academic Press, London, 1st edn., ISBN 0-12-058575-8, 1987.
- Baker, D. J. and Stair Jr., A. T.: Rocket measurements of the altitude distributions of the hydroxyl airglow, *Physica Scripta*, 37, 611, <https://doi.org/10.1088/0031-8949/37/4/021>, 1988.
- 270 Beig, G.: Long term trends in the temperature of the mesosphere/lower thermosphere region: 1. Anthropogenic influences, *Journal of Geophysical Research*, 116, A00H11, <https://doi.org/10.1029/2011JA016646>, 2011a.
- Beig, G.: Long-term trends in the temperature of the mesosphere/lower thermosphere region: 2. Solar response, *Journal of Geophysical Research: Space Physics*, 116, <https://doi.org/10.1029/2011JA016766>, 2011b.
- Bittner, M., Offermann, D., and Graef, H.-H.: Mesopause temperature variability above a midlatitude station in Europe, *Journal of Geophysical Research: Atmospheres*, 105, 2045–2058, <https://doi.org/10.1029/1999JD900307>, 2000.
- 275 Bittner, M., Offermann, D., Graef, H.-H., Donner, M., and Hamilton, K.: An 18-year time series of OH rotational temperatures and middle atmosphere decadal variations, *Journal of Atmospheric and Solar-Terrestrial Physics*, 64, 1147–1166, [https://doi.org/10.1016/S1364-6826\(02\)00065-2](https://doi.org/10.1016/S1364-6826(02)00065-2), 2002.
- French, W. J. R. and Mulligan, F. J.: Stability of temperatures from TIMED/SABER v1.07 (2002–2009) and Aura/MLS v2.2 (2004–2009) compared with OH(6–2) temperatures observed at Davis Station, Antarctica, *Atmospheric Chemistry and Physics*, 10, 11 439–11 446, <https://doi.org/10.5194/acp-10-11439-2010>, 2010.
- 280 French, W. J. R., Mulligan, F. J., and Klekociuk, A. R.: Analysis of 24 years of mesopause region OH rotational temperature observations at Davis, Antarctica – Part 1: long-term trends, *Atmospheric Chemistry and Physics*, 20, 6379–6394, <https://doi.org/10.5194/acp-20-6379-2020>, 2020.
- 285 Fritts, D. C. and Alexander, M. J.: Gravity wave dynamics and effects in the middle atmosphere, *Reviews of Geophysics*, 41, <https://doi.org/10.1029/2001RG000106>, 2003.
- García-Comas, M., López-González, M. J., González-Galindo, F., de la Rosa, J. L., López-Puertas, M., Shepherd, M. G., and Shepherd, G. G.: Mesospheric OH layer altitude at midlatitudes: variability over the Sierra Nevada Observatory in Granada, Spain (37° N, 3° W), *Annales Geophysicae*, 35, 1151–1164, <https://doi.org/10.5194/angeo-35-1151-2017>, 2017.
- 290 Holmen, S. E., Dyrland, M. E., and Sigernes, F.: Mesospheric temperatures derived from three decades of hydroxyl airglow measurements from Longyearbyen, Svalbard (78°N), *Acta Geophysica*, 62, 302–315, <https://doi.org/10.2478/s11600-013-0159-4>, 2014.
- Holton, J. R.: The Influence of Gravity Wave Breaking on the General Circulation of the Middle Atmosphere, *Journal of Atmospheric Sciences*, 40, 2497 – 2507, [https://doi.org/10.1175/1520-0469\(1983\)040<2497:TIOGWB>2.0.CO;2](https://doi.org/10.1175/1520-0469(1983)040<2497:TIOGWB>2.0.CO;2), 1983.
- Kalicsinsky, C., Knieling, P., Koppmann, R., Offermann, D., Steinbrecht, W., and Wintel, J.: Long-term dynamics of OH* temperatures over central Europe: trends and solar correlations, *Atmospheric Chemistry and Physics*, 16, 15 033–15 047, <https://doi.org/10.5194/acp-16-15033-2016>, 2016.
- 295 Kalicsinsky, C., Kirchhoff, S., Knieling, P., and Zlotos, L. O.: Long-term variations in the mesopause region derived from OH*(3,1) rotational temperature observations at Wuppertal, Germany, from 1988 – 2022, *Advances in Space Research*, 73, 3398–3407, <https://doi.org/10.1016/j.asr.2023.08.045>, recent advances in equatorial, low- and mid-latitude mesosphere, thermosphere and ionosphere studies, 2024.
- 300 Kim, G., Kim, J.-H., Kim, Y. H., and Lee, Y. S.: Long-term trend of mesospheric temperatures over Kiruna (68°N, 21°E) during 2003–2014, *Journal of Atmospheric and Solar-Terrestrial Physics*, 161, 83–87, <https://doi.org/10.1016/j.jastp.2017.06.018>, 2017.



- Krassovsky, V., Shefov, N., and Yarin, V.: Atlas of the airglow spectrum 3000–12400 Å, *Planetary and Space Science*, 9, 883–915, [https://doi.org/https://doi.org/10.1016/0032-0633\(62\)90008-9](https://doi.org/https://doi.org/10.1016/0032-0633(62)90008-9), 1962.
- 305 Lange, G.: Messung der Infrarotemissionen von OH* und O₂ in der Mesosphäre, Ph.d. thesis, University of Wuppertal, Wuppertal, 1982.
- Meinel, A. B.: OH Emission Bands in the Spectrum of the Night Sky. I, *The Astrophysical Journal*, 111, 555, <https://doi.org/10.1086/145296>, 1950.
- Mies, F. H.: Calculated vibrational transition probabilities of OH(X²I), *Journal of Molecular Spectroscopy*, 53, 150–188, [https://doi.org/https://doi.org/10.1016/0022-2852\(74\)90125-8](https://doi.org/https://doi.org/10.1016/0022-2852(74)90125-8), 1974.
- 310 Mulligan, F. J. and Lowe, R. P.: OH-equivalent temperatures derived from ACE-FTS and SABER temperature profiles - a comparison with OH*(3-1) temperatures from Maynooth (53.2°; N, 6.4°; W), *Annales Geophysicae*, 26, 795–811, <https://doi.org/10.5194/angeo-26-795-2008>, 2008.
- Oberheide, J., Offermann, D., Russell III, J. M., and Mlynczak, M. G.: Intercomparison of kinetic temperature from 15 μm CO₂ limb emissions and OH*(3,1) rotational temperature in nearly coincident air masses: SABER, GRIPS, *Geophysical Research Letters*, 33, <https://doi.org/https://doi.org/10.1029/2006GL026439>, 2006.
- 315 Offermann, D., Wintel, J., Kalicinsky, C., Knieling, P., Koppmann, R., and Steinbrecht, W.: Long-term development of short-period gravity waves in middle Europe, *Journal of Geophysical Research: Atmospheres*, 116, D00P07, <https://doi.org/10.1029/2010JD015544>, 2011.
- Perminov, V., Semenov, A., Medvedeva, I., and Zheleznov, Y.: Variability of mesopause temperature from the hydroxyl airglow observations over mid-latitudinal sites, Zvenigorod and Tory, Russia, *Advances in Space Research*, 54, 2511–2517, <https://doi.org/https://doi.org/10.1016/j.asr.2014.01.027>, forcing on the Atmosphere, 2014.
- 320 Schmidt, C., Höppner, K., and Bittner, M.: A ground-based spectrometer equipped with an InGaAs array for routine observations of OH(3-1) rotational temperatures in the mesopause region, *Journal of Atmospheric and Solar-Terrestrial Physics*, 102, 125–139, <https://doi.org/10.1016/j.jastp.2013.05.001>, 2013.
- Sivjee, G. G. and Hamwey, R. M.: Temperature and chemistry of the polar mesopause OH, *Journal of Geophysical Research: Space Physics*, 92, 4663–4672, <https://doi.org/https://doi.org/10.1029/JA092iA05p04663>, 1987.
- 325 Wüst, S., Bittner, M., Espy, P. J., French, W. J. R., and Mulligan, F. J.: Hydroxyl airglow observations for investigating atmospheric dynamics: results and challenges, *Atmospheric Chemistry and Physics*, 23, 1599–1618, <https://doi.org/10.5194/acp-23-1599-2023>, 2023.