Review report for Ballington and Kalicinsky: "Physics-Informed Deep Learning for OH(3–1) Airglow Spectra Temperature Retrieval"
https://doi.org/10.5194/egusphere-2026-4868

The authors present a new approach of determining rotational temperatures from OH airglow spectra by applying deep learning algorithms to the raw data. Their work is essentially divided into two parts: 1) the application of the aforementioned AI methods to classify the observational data, thereby filtering out spectra that are not at all suitable for further processing (e.g. due to cloud coverage), and 2) the subsequent further processing using a so-called physics-informed AI model to determine the rotational temperatures.

From my point of view, the first point is quite clear, and I personally believe it is long overdue for someone in the airglow community to finally address this issue. I would therefore like to offer my sincere congratulations to the authors on accomplishing this task.

However, I don't fully understand the point of the second approach, to put it rather bluntly: "Why put an AI model through its paces for a calculation that is, in principle, pure analysis?" (On the other hand: I'm a dinosaur and scared of new things) Probably, there are good reasons, but the authors discuss this point rather briefly. I therefore identify two main points that should be explored in greater depth (my major comments). In addition, there are a number of minor points that require a small revision or further justification.

All in all, once the necessary revisions have been made, the paper will make an important contribution to the field through the (in my view, long overdue) introduction of AI methods into airglow spectroscopy.

Since I think it will take some effort to address my major comments (especially #1: a significantly more detailed comparison of the final results of the established method and the new method is of uttermost importance!), I suggest a major revision to provide enough time for the revisions to be made.

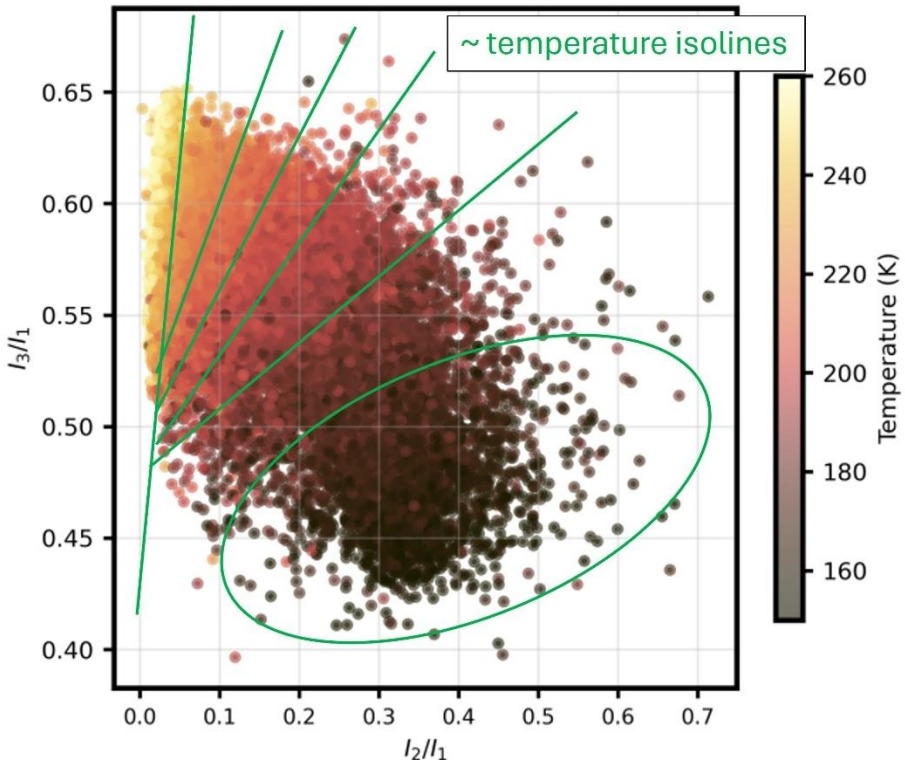
Major comments

1. *Figure 8 / L 226/227: For a fair comparison, the nightly means shown in Fig. 8 are computed only from spectra which both methods classified as valid*.

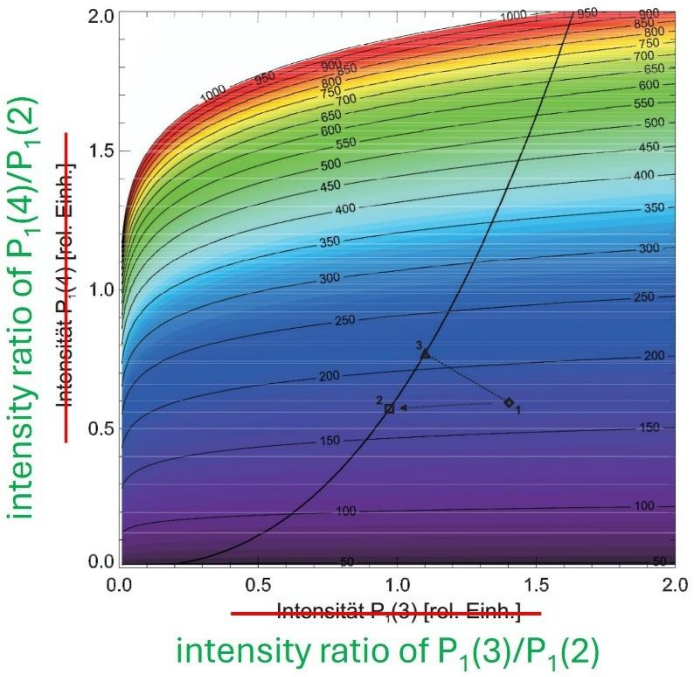
   Nice try! But a fair comparison means you present both, 1) the results obtained from the cases identified as good by both methods (which you did and which allows to calculate the differences between the two methods), and 2) the results obtained from each method individually. Afterall, if I understand correctly, you are trying to establish the new method as a replacement for the old method that still requires a lot of manual work. That's reasonable. But therefore you need to show one year (or more years) processed by method A "as is" and compare it to "as it will be in the future".
   From Figure 1 it is rather clear that the established method yields nightly means above 230K in any given year and temperatures below 165K in ~7 years. But Figure 8 shows no nightly means above 230K and only five values between 165K and 170K. Either, this is due to the year 2025 being rather special or it's due to the ~20% false negatives (see confusion matrix) making the methods differ.
   That doesn't necessarily mean that the new method performs worse, but clearly there are differences that need to be addressed. Under which circumstances does the new method yield nightly mean values when the old doesn't? - Maybe never due to the tuning to avoid false positives) - When does the old method yield temperatures when the new method doesn't? (Highly likely, as only 81.6% of the old good spectra are recognized as such by the new method. Is there a bias towards more 'warm' (>230K) nights the missing? Are these nights characterized by bad weather / bad data? Is the new method performing better here or is it too restrictive?
   Judging from Figure 1: is the training data biased towards 190K-230K and avoiding low temperatures?

   l.34/35: "*Prior to GRIPS-N, the requirement was one hour for high-quality spectra or two hours otherwise (Bittner et al., 2002).*"
   What are the thresholds for your nightly means today? Are there any at all? (I vaguely remember Schmidt et al. (2023), https://amt.copernicus.org/articles/16/4331/2023/ trying to overcome the issue of rather arbitrary one hour-, two hour-, four hour-, … thresholds by introducing a new approach). But I cannot identify your approach.
   Out of curiosity, I looked up Bittner's definition of "high-quality" and "otherwise". As it turns out "high-quality" equals an SNR ~25 and "normal" equals an SNR of ~15. You on the other hand work with an SNR of ~6. How does this compare to the statement that you try to ensure "data continuity" (… not to ask of how to actually calculate / define the SNR in these spectra):
   l. 29-32: "*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*."

2. I am a bit unhappy with Figure 6 – happy that you provide it, but irritated by how it looks like when it comes to the details and unhappy because I_2/I_1 & I_3/I_1 are not the actual peak intensity ratios, which make them hard to judge for anybody incl. ~~the reviewers~~ (at least one reviewer). Your reasoning is plausible, but plausibility does not mean that the result is correct (I am old enough to remember the first calculators producing at times bizarre results – like sqrt(4) = 1.999999999, 'almost there but not entirely').
   Let me therefore explain what looks irritating to me in this context: I_2/I_1 shall serve as proxies for the true peak intensity ratios $P_1(3)/P_1(2)$. In Figure 2 we can see that $P_1(3)/P_1(2)$ is approximately 1.0 and $P_1(4)/P_1(2)$ is approximately 0.8. From experience I would state that $P_1(3)/P_1(2)$ typically ranges from 0.8 to 1.2 and $P_1(4)/P_1(2)$ ranges from 0.5 to 0.7. Despite the absolute numbers, the former is (almost) always larger than the latter (unless T >> 300K). Here, we see the opposite: I_3/I_1 ~ 0.45-0.65, but I_2/I_1 ~0.05-0.5. Now I am irritated, because of what the plot looks like. Everything might still be somehow correct, but it's not the actual intensity ratios and it's not "approximately" the intensity ratios; at the moment I am unable to judge what it actually is. It is "something else", sort of a black box producing a pattern that allows for deriving meaningful temperatures.



   Note: potential isolines (green) run almost vertically. For comparison, in the depths of the internet I found a graph showing the 'true' distribution (Figure 2.11 from: https://elib.dlr.de/108415/1/Schmidt_Carsten_FB-2016-49.pdf, "Schmidt (2016)", last access 2026-11-24) and thanks to auto translation (it's written in German) it is more or less clear (even to me) how it was constructed.

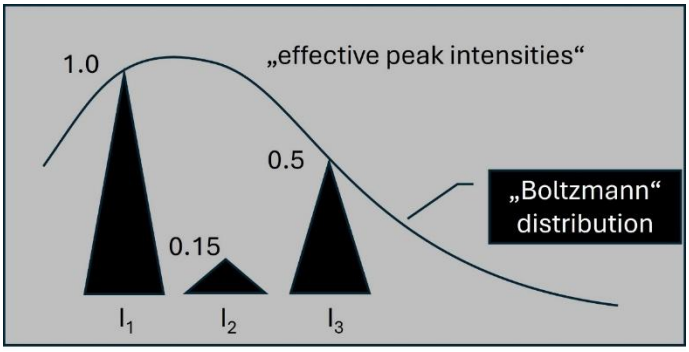


Here, the thick line refers to the forward solution, i.e. „intensity as a function of (given) temperature" (aka "the Boltzmann relationship"), with the coloured space representing the potential observational space, i.e. any possible intensity distribution.

What I find interesting: the isolines (and colours) represent the rotational temperatures resulting from the given intensity distribution. In the region of 'benign' spectra, they tend to run more or less horizontally in this plot.

It is only in the extreme range (small $P_1(3)$, which is the middle peak in the spectrum) that they become vertical. I hope this shows that I_2/I_1 & I_3/I_1 cannot resemble "effective peak ratios" $P_1(3)/P_1(2)$ & $P_1(4)/P_1(2)$ and it's not as 'simple' as stated in *lines 156/157: "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."*

By the way: this is how the "*effective peak ratios*" would look like (which is anything but a thermalized distribution regardless of the actual meaning… sort of a population inversion: more molecules in states "1" and "3" than "2"). Long story short: at the moment and with the information provided I cannot relate the so-called "effective peak ratios" to a real physical feature. This needs further discussion.



Minor comments

1. Throughout the entire manuscript: please consider using $P_1(2)$, $P_1(3)$, $P_1(4)$ instead of ~~P1(2), P1(3), P1(4)~~; applies of course also to the P_2-lines. It is common practice to distinguish between the two branches using subscript 1 or 2.

2. The authors emphasize the importance of continuity concerning the construction of long-term time series. Yet, they train a neural network only for the new (or overhauled GRIPS-II?) instrument, which is in operation since 2014: is it possible that the new approach required such refined training that it isn't readily applicable to the older data and performs worse than the previous scientists doing it manually?
   In your defense: you developed something new and it's absolutely fair to start with one instrument only, but at the beginning of the manuscript I got the impression that doing it for the long Wuppertal time series is what you were up to.

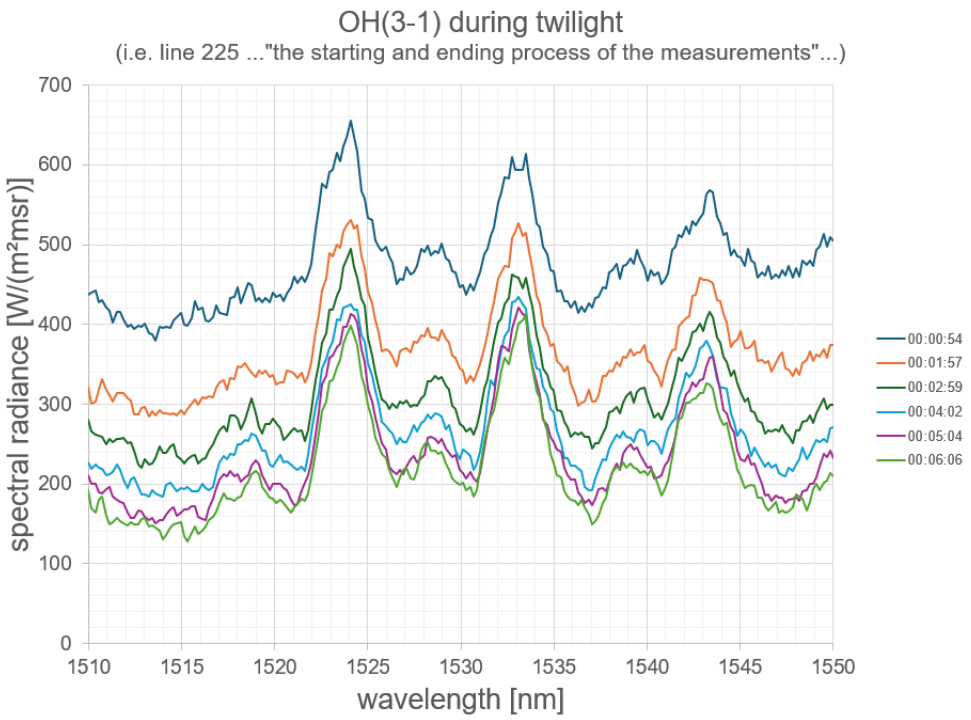
3. Section 3.3, l122/123: I am by no means an expert on AI methods, but I vaguely remember having read a convolutional layer would always use more than on kernel. Please verify: … *Convolutional layers apply a learned kernel… (a set of kernels…?)*

4. L194/195: *For comparison, an unconstrained dense baseline model (~ 681000 parameters, no analytical Boltzmann layer) was trained on the same data. The Boltzmann model achieves slightly superior mean performance with less than 10% parameters.*

   Thank you for providing details on many other aspects of your work. But here, the description is a bit short: "slightly superior mean performance" expressed by which parameters? Precision? Confusion matrix? RMSE? (…)? How much better quantitatively? A hint that 10% refer to the above-mentioned 50 000 parameters would be welcome.

5. L. 224/226: *… 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 the recursive correction for the P1(4) overlap does not converge.*

   a. As far as I understand "*the starting and ending process of the measurements":* the start/end of the night is the issue here. This means twilight, and twilight causes neither noise nor fluctuations but a systematic change of background radiances. I attach a set of twilight spectra acquired with a non-scanning spectrometer and separated in time by ~1 minute. A scanning spectrometer like GRIPS-N will be highly affected by the changing background. This will be the reason for the legacy pipeline to consequently exclude these spectra. Concerning the new predictions: clearly, there are 3 lines in all spectra, but if there are no sufficient training data for twilight conditions, then the new algorithm will encounter issues dealing with them.
      On the other hand: if all of these spectra (100%) are flagged accordingly in the training data (because that's what the legacy pipeline does), then why are they not recognized as such?



b. Isn't it a clear hint, that something's wrong with the spectrum if the recursive correction doesn't converge? Having looked at Lange (1982) and Schmidt (et al., 2013; 2016 (the latter, 2016, also being the source of the Figure in my major comment #2)) this iteration appears to be a straightforward approach. If it doesn't converge, then the spectrum should be odd, i.e. be an intensity distribution, that isn't thermalized. How often does the new approach identify such a spectrum as good? Is this number within the boundaries provided by the confusion matrix?

6. Referencing to Schmidt et al. (2013) at a couple of instances in section 2.2 concerning the details of the retrieval is a bit misleading in its current formulation. Schmidt et al. (2013) did not actually develop or change details of the retrieval, instead they present a short summary of the works they rely on, emphasizing that a couple of the studies they reference to are not easily accessible (e.g. Kvifte, 1959 or Lange, 1982), which is especially true for the latter as I can now confirm.
   Therefore, I suggest to rephrase:
   l72: … constant (Meinel, 1950; see also discussion in Schmidt et al., 2013).
   l75: … unknowns (Lange, 1982; Schmidt et al. 2013 and references therein).

Spelling / grammar

This long list looks worse than it is. The manuscript is written rather well and I am glad to provide some corrections and a couple of suggestions for further improvements.

l18/19, *'surface' might be a bit misleading, troposphere (and maybe stratosphere) fits better here*: These waves are crucial for transporting energy from the ~~surface~~ troposphere and stratosphere to the upper…

l22: … network (e.g. Bittner…

l26/27: … today (e.g. Bittner…

l39/40: *since YOU are able to develop an efficient method in line 38, YOU can also construct a neural network and avoid the passive construction in lines 39/40*: In this work, ~~an algorithm based on deep neural networks is constructed~~ we construct an algorithm based on deep neural networks…

l40/41 *grammar, one cannot 'allow a network'; pay attention to keep the hyphen in 'custom-tailored' (was deleted during copy-pasting into this document in my case)*: … allowing for automatic classification and filtering of OH spectra, as well as introducing/establishing a custom-tailored network that…

l42, suggest to rephrase: This work ~~touches on~~ highlights/demonstrates both the merit and importance of

l55/56 *suggestion to rephrase*: The instrument has collected data from Wuppertal (51◦ N, 7◦ E) since 2014~~ after~~ , following the instrument reconstruction and ~~the~~ final set-up of GRIPS-N…

l56: … the complete time~~s~~ series of…

l58: As with ~~real~~ any measurement data…

l58/59: Data gaps arise primarily due to cloud cover, ~~which is~~ the dominant cause of missing observations…

l59/60 *suggestion to rephrase*: **Contamination** from artificial light sources, including a mercury (Hg) line at 1529.6 nm that ~~**contaminates**~~ overlaps the P~~1~~$_1$(3) emission

l60: … must also be accounted for, and…

l62 *suggestion to rephrase*: …fluctuations on ~~the~~ timescales of ~~several~~ a few minutes…

l65 *suggestion to rephrase*: The following section provides… / ~~The following~~ This section provides…

l72/74 *both are correct, but interchanging them in the same paragraph is a bit inconsistent*: Q = "state sum" vs. Q = "partition sum"

l74: In practice, the formula…

l82/83: Since… OH(3-1) P-branch, an iterative correction is…

l92: The raw dataset ~~is comprised of~~ comprises 11 years…

l98: … these cases can be safely ~~pruned~~ discarded.

l99: … characteristics, ~~as~~ for example in early…

l102: … The nights are randomly ~~shuffled~~ distributed across the sets to avoid…

l103/104: … classified as good, ~~thus~~ i.e., nights with…

l105: … effective ~~splits~~ split ratios for…

l112: … whereas in the test set it is approximately 60/40…

l112/113: Example good- and bad-labelled…

l116: … may not be suitable…

l115: Figure 2: Example good- and bad-labelled intensity spectra.

l119: … [0,1], thereby removing…

l128: Each branch ~~is comprised of~~ comprises two…

l133: Is there spatial information or is it spectral information? … the **spatial** information…

l138: *change from previous present tense*: The hyperparameters **were** determined…

l145: … is replaced ~~with~~ by a neural network encoder

l146: … to yield the temperature.

l149: … output values do not correspond…

l149/150: … produce a ~~real~~ (physically) meaningful predicted temperature.

l153: … this ~~effect~~ issue is avoided entirely.

l183: These ~~borderline~~ intermediate cases…

l187: … the model ~~is able~~ to predicts extreme temperatures with ~~(very)~~ little ~~visible~~ apparent systematic bias.

l188/189: The model **has a** mean **bias**… , **standard deviation**… , and ~~the~~ has **a 90th percentile error** ~~is~~ of 9.40 K.

l190-192: The standard deviation ~~for~~ of 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~~ quiet-time intervals ***or*** periods up to more than 30 K ~~in~~ during times with large…

l193/194: …noise uncertainty in ~~quite~~ quiet time intervals.

l196: This implies a significantly ~~lighter~~ reduced training cost,

l196: *change of tense*: … and ~~the model~~ it ~~showed~~ shows more stable…

l205: … plotted as ~~abstract~~ effective values …

l206/207: … consistent with the theory ~~of~~ discussed in Sect. 2.2

l210: … on both a per-spectrum and a per-night basis.

l210: … shows the per-spectrum concordance ~~scatter~~ plot (*or* scatter plot, *but not both*), …

l212: ~~The directional prediction of the model is accurate to 91.5%~~… The model achieves a directional accuracy of 91.5%…

l214: … direction in 93.8% of ~~the time~~ cases.

l218: … and a minimum in ~~the~~ summer. (*or* … ~~in~~ during the summer.)

l219: This ~~**uncertainty**~~ is far below the statistical **uncertainty** of a typical nightly mean temperature ~~for~~ from the legacy pipeline, which…

l220: … of ~~single~~ individual measurements…

l223/224: … in some cases, the first or last spectra of the night are ~~taken from~~ accepted by the network but are always excluded ~~in~~ by the legacy pipeline…

l224/225 (*see also respective minor comment*): … may contain fluctuations ~~or~~, noise or systematic biases ~~because of the starting and ending process of the measurements~~ during twilight conditions *(or* near sunrise/sunset *or* civil and nautical twilight*)*, …

l227/228: ~~It is visually clear~~ As is evident, especially from the residual plot, ~~that~~ the …

l230/231: … the uncertainty ~~in~~ of the retrieved temperatures relative to those from the legacy pipeline…

l231/232: The presented framework ~~used has general applicability~~ is broadly applicable to a…

l234: The GRIPS-N instrument ~~underpins~~ contributes to one of the …

l240: … distributed across the ~~full~~ entire observation period …