Airglow data-driven modeling over a period of three solar cycles

Simon Mackovjak[1], Matej Varga[2], Stanislav Hrivňak[3], Ondrej Palkoci[3], and Goderdzi Didebulidze[4]

[1]Institute of Experimental Physics
[2]Faculty of Electrical Engineering and Informatics, Technical University of Kosice
[3]GlobalLogic s.r.o.
[4]Ilia State University

November 30, 2022

Abstract

The Earth's upper atmosphere is a dynamic environment that is continuously affected by space weather from above and atmospheric processes from below. An effective way to observe this interface region is the monitoring of airglow. Since the 1950s, airglow emissions have been systematically measured by ground-based photometers in specific wavelength bands during the nighttime. The availability of the calibrated data from over 30 years of photometric airglow measurements at Abastumani in Georgia (41.75 N, 42.82 E), at wavelengths of 557.7 nm and 630.0 nm, enable us to investigate if a data-driven model based on advanced machine learning techniques can be successfully employed for modeling airglow intensities. A regression task was performed using the time series of space weather indices and thermosphere-ionosphere parameters. We have found that the developed data-driven model has good consistency with the commonly used GLOW airglow model and also captures airglow variations caused by cycles of solar activity and changes of the seasons. This enables us to visualize the green and red airglow variations over a period of three solar cycles with a one-hour time resolution.

Posted on 30 Nov 2022 | CC-BY-NC 4 | https://doi.org/10.1002/essoar.10505187/v1 | This a preprint and has not been peer reviewed. Data may be preliminary.

Data-driven modeling of atomic oxygen airglow over a period of three solar cycles

Š. Mackovjak[1,3], M. Varga[2], S. Hrivňak[3], O. Palkoci[3], G. G. Didebulidze[4]

[1]Department of Space Physics, Institute of Experimental Physics, Slovak Academy of Sciences, Košice, Slovakia

[2]Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Košice, Slovakia

[3]GlobalLogic Slovakia s.r.o., Košice, Slovakia

[4]Georgian National Astrophysical Observatory, Ilia State University, Tbilisi, Georgia

Key Points:

- A data-driven model is able to represent complex physical phenomena
- Advanced machine learning techniques are effective for the development of the data-driven model
- Developed data-driven model visualizes airglow hourly intensities over a 30-year period for location Abastumani (41.75° N, 42.82° E)

Corresponding author: Šimon Mackovjak, Institute of Experimental Physics SAS, Watsonova 47, 04001 Košice, Slovakia, `mackovjak@saske.sk`

Abstract

The Earth's upper atmosphere is a dynamic environment that is continuously affected by space weather from above and atmospheric processes from below. An effective way to observe this interface region is the monitoring of airglow. Since the 1950s, airglow emissions have been systematically measured by ground-based photometers in specific wavelength bands during the nighttime. The availability of the calibrated data from over 30 years of photometric airglow measurements at Abastumani in Georgia (41.75° N, 42.82° E), at wavelengths of 557.7 nm and 630.0 nm, enable us to investigate if a data-driven model based on advanced machine learning techniques can be successfully employed for modeling airglow intensities. A regression task was performed using the time series of space weather indices and thermosphere-ionosphere parameters. We have found that the developed data-driven model has good consistency with the commonly used GLOW airglow model and also captures airglow variations caused by cycles of solar activity and changes of the seasons. This enables us to visualize the green and red airglow variations over a period of three solar cycles with a one-hour time resolution.

1 Introduction

The Earth's upper atmosphere acts as an interface between processes in space and on Earth. It is a very dynamic environment continuously influenced by solar radiation and space weather from above and by atmospheric weather and electrical discharges from below (Pfaff, 2012). An effective way to monitor these dynamics during night-time periods in the altitude range of 80–300 km is observation of airglow (Khomich et al., 2008). Airglow is a non-thermal emission of light originating from excited atomic or molecular states. The source of the excitation, directly or indirectly, is the solar electromagnetic radiation (von Savigny, 2017). The particular process responsible for the emission of airglow and the amount of this emission is mainly dependent on the composition and concentrations of neutral constituents and ion/electron densities in the thermosphere-ionosphere system.

The earliest reported airglow variation is connected to the 11-year long solar cycle. The correlation between the well-known atomic oxygen OI($^1D_2 - {}^1S_0$) airglow emission of the green line (557.7 nm) with sunspot area was revealed in 1935 (Rayleigh & Jones, 1935). The connection of solar activity, expressed by solar flux index F10.7 was confirmed by extensive studies (Deutsch & Hernandez, 2003; Liu & Shepherd, 2008; Reid et al., 2014).

The authors provide clear evidence that the green line intensity is maximal during the maximum of the solar cycle. The variations within the year (annual oscillation and semi-annual oscillation) are associated with the yearly tilt and rotation of the Earth around the Sun and also with the dynamics in the whole atmosphere, mainly driven by atmospheric tides. The amplitude and maximum of a period are different for different locations. Shepherd et al. (2006) and Liu et al. (2008) used UARS/WINDII (Shepherd et al., 1993) space-based observations of the green line in the years 1991–1997 to present airglow variations during the year for different latitudes. The authors concluded that for the equatorial region, semi-annual variation has maxima at equinoxes and for the mid-latitude regions, the annual variation is dominant and has a maximum in autumn in the northern hemisphere and in spring in the southern hemisphere. There are also shorter and non-periodic variations in the upper atmosphere. The influence of geomagnetic storms has been observed in airglow intensity measurements since the mid-twentieth century (Silverman, 1970). During a geomagnetic storm, the density distribution of the ions and neutral constituents in the upper atmosphere varies dramatically. Such variations may have signatures in airglow emissions (Leonovich et al., 2011; Makela et al., 2014; Bag et al., 2017).

Although some patterns in airglow variations were recognized, a clear physical explanation is still missing. This is a consequence of the very high complexity of the environment and the fact that the response of airglow production might be not uniformly related to a single process. Indeed, airglow intensity represents the continuous variation of solar activity, solar wind, interplanetary magnetic field, magnetospheric drivers as well as non-constant density and temperature conditions in the upper atmosphere together with ever-present vertical motions from lower atmosphere including tides, planetary waves, and atmospheric gravity waves. The ionosphere-thermosphere system is also affected by alteration of the global ionosphere electric potential and by various ionospheric instabilities such as plasma bubbles and ionospheric scintillation (Eastes et al., 2019). As the understanding of consequences of these processes is still not sufficient, the whole subject is very topical and it is an objective of several ground-based and space-based missions (e.g. Eastes et al., 2017; Immel et al., 2018; Hannawald et al., 2019; Mackovjak et al., 2019; Wüst et al., 2019).

Data-driven machine learning techniques have become effective tools in space science in recent years (e.g. Ball & Brunner, 2010; George & Huerta, 2018; Zucker & Giryes, 2018). It is mainly due to the fact that the huge amount of space data can be effectively processed by powerful computing units utilizing open source frameworks supported by technology giants (e.g. Pedregosa et al., 2011; Abadi et al., 2015; Paszke et al., 2017). A comprehensive overview of the machine learning techniques and their application for space weather research is presented by Camporeale et al. (2018). The aim of this paper is to investigate if a data-driven approach using machine learning techniques can provide adequate results of long-term airglow intensity modeling. The usefulness of this approach will be evaluated by its capability to reproduce generally known airglow variations as well as by comparison with the output from the Global Airglow (GLOW) model (Solomon et al., 1988; Solomon, 2017; Hirsch & Solomon, 2019). The data and machine learning methods used are described in Section 2. The results obtained and discussion are presented together in Section 3. Section 4 summarizes our work and describes the next steps in our research.

2 Data and Methods

Depending on the solar elevation, airglow can be categorized as dayglow, twilightglow and nightglow (von Savigny, 2017). Dayglow emission is the brightest but its observation is not straightforward due to the presence of direct and scattered light from the Sun. Therefore, every time the term airglow is used in this work, the nightglow (solar zenith angle (S_{ZA}) is higher than 108°) is considered. Our focus is on atomic oxygen emissions - green line and red line with the wavelengths 557.7 nm and 630.0 nm, respectively. The details of their emission production mechanisms are presented in Khomich et al. (2008).

The main dataset used consists of calibrated photometric data of the airglow green line (557.7 nm) and airglow red line (630.0 nm) measured at Abastumani in Georgia (41.75° N, 42.82° E, 1,580 m above sea level) in the years 1957–1993 (Fishkova, 1983; Gudadze et al., 2007; Didebulidze et al., 2011; NDMC, (last access: November 30, 2020)). Measured intensities are in units of Rayleighs ($1\,\mathrm{R} = 10^{10}$ photon m^{-2} s^{-1}). They were acquired during the moonless (moon zenith angle (M_{ZA}) is higher than 90°) and cloudless conditions. The time resolution of the data is 6–15 minutes. For the purposes of this work, hourly averages were used within the time interval 1964–1993. The boxplots of the measured data are displayed in Figure 1. They represent the distributions of the measurements over the years. The total amount of data used is $\sim$3,850 measurements, rep-

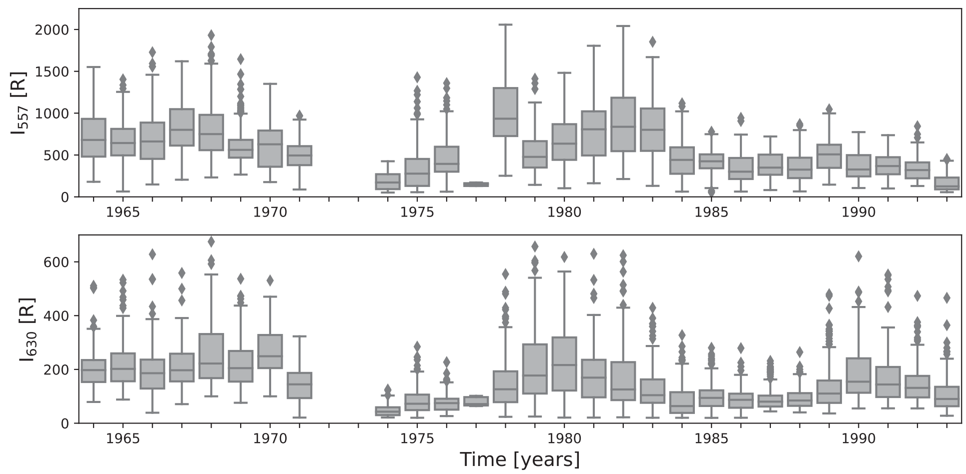


Figure 1. The box plots of the airglow measurements at Abastumani (Georgia) over the years 1964–1993. Only the hourly averages are considered where the sunless, moonless, and cloudless conditions are satisfied. Each interquartile range is represented by the particular box and the median of the distribution is marked with a horizontal dash. The diamond points outside the box whiskers represent the outliers caused by high variability of the measurements. They are not caused by an experimental error and they can be used in the analysis.
Distributions of the green line and red line intensities are displayed on the top and bottom, respectively.

resenting $\sim 8\,\%$ of all possible dark night hours (hours when $S_{ZA} > 108°$ & $M_{ZA} > 90°$) over a 30-year period for this location. One of the goals of this work is to model the airglow green and red line intensities for the rest of the dark night hours (i.e. $\sim 92\,\%$) in this period.

In the data-driven modeling approach, the measured airglow intensities were used as labels (target outputs). The features (inputs) for the model were chosen from four categories: space weather indices, thermosphere parameters, ionosphere parameters, and Sun-Earth distance. These four categories cover the basic processes that affect airglow intensities. Although the exact physical relations between these features (inputs) and labels (target outputs) are not considered here, it is assumed that these underlying relations are present in the data. Machine learning techniques should be able to recognize these underlying relations and also model airglow intensities for previously unseen feature values. For the appropriate feature selection, all available data from the OMNIWeb

Table 1. The selected features for machine learning techniques to model airglow intensities

Feature	Units	Description	Source
F10.7 index	SFU	Solar radio flux per frequency ($\lambda = 10.7\,\mathrm{cm}$)	OMNIWeb[a]
Kp index		Geomagnetic planetary K-index	OMNIWeb[a]
Dst index	nT	Geomagnetic equatorial index	OMNIWeb[a]
Neutral Temperature	K	Temperature of neutral atmosphere	NRLMSISE-00[b]
Total Mass Density	g/cm^3	Total mass density of neutral atmosphere	NRLMSISE-00[b]
O	N/cm^3	Atomic oxygen density	NRLMSISE-00[b]
O_2	N/cm^3	Molecular oxygen density	NRLMSISE-00[b]
N	N/cm^3	Atomic nitrogen density	NRLMSISE-00[b]
N_2	N/cm^3	Molecular nitrogen density	NRLMSISE-00[b]
H	N/cm^3	Atomic hydrogen density	NRLMSISE-00[b]
T_e	K	Temperature of electrons	IRI-2016[c]
n_e	N/m^3	Density of electrons	IRI-2016[c]
h_mF_2	km	F_2 layer peak height	IRI-2016[c]
N_mF_2	N/m^3	F_2 layer peak density	IRI-2016[c]
Sun-Earth	AU	Sun-Earth distance	PyEphem[d]

[a]Available at: https://omniweb.gsfc.nasa.gov/form/dx1.html (King & Papitashvili, 2005)

[b]Available at: https://ccmc.gsfc.nasa.gov/modelweb/models/nrlmsise00.php (Picone et al., 2002)

[c]Available at: https://ccmc.gsfc.nasa.gov/modelweb/models/iri2016_vitmo.php (Bilitza et al., 2017)

[d]Available at: https://pypi.org/project/pyephem

space weather database (King & Papitashvili, 2005), NRLMSISE-00 thermosphere model (Picone et al., 2002), and IRI-2016 ionosphere model (Bilitza et al., 2017) were explored. These data are accessible in hourly resolution. The availability of the features for a 30-year interval was considered in the feature selection process. The parameters of the neutral atmospheres and ionosphere are obtained for the nominal altitudes 95 km and 250 km for modeling green and red line, respectively. These are the altitudes of particular peak airglow layer emissions (von Savigny, 2017). The feature selection was mainly guided by current physical understanding of the features' influence on airglow production and also on automatic data characterization methods. Automatic methods such as univariate feature selection and recursive selection of the features based on the model training process (Pedregosa et al., 2011) have been examined for the exclusion of the redundant features by quantification of their mutual correlation and by other statistical tests. The list of 15 features selected for our work is presented in Table 1. We would like to mention that none of the investigated features had a significant correlation with airglow intensities. The absolute value of pairwise Pearson correlation coefficient was not higher than 0.26 for any pair of feature and label. It is noted that consideration of additional features did not lead to better results. This does not mean the irrelevance of other indices such as e.g. the interplanetary magnetic field or solar wind parameters. These indices were excluded as their availability is less than 60% of the studied time interval.

The modeling of airglow intensities using the space weather indices, thermosphere-ionosphere parameters, and Sun-Earth distances as the input is indeed a regression problem. Using known input and output values we would like to approximate the mapping function that could provide, with sufficient precision, the airglow intensities as the output for the previously unseen inputs. In the current work, we have employed the following supervised machine learning techniques for the regression problem: linear regression, Neural Networks, and the ensemble algorithms - Random Forest and Extreme Gradient Boosting (XGBoost). Ordinary least squares linear regression, as the common statistical approach in astronomy (Isobe et al., 1990), was used as the simplest technique. The Neural Network is one of the most popular machine learning techniques, although its usage is not always the best option, especially for problems where the features come from different distributions (Fernández-Delgado et al., 2014). It is based on the fact that every continuous real function over a compact set of real numbers can be approximated arbitrarily well by a function defined as a Neural Network with a high enough number of neurons. For more details refer to Cybenko (1989). In this work, we used a single hidden layer feed-forward Neural Network with a number of neurons 128-128-1 (i.e. 128 neurons in the input layer, 128 neurons in the hidden layer, 1 neurons in the output layer), hyperbolic tangent activation function, 300 learning epochs, and learning rate from 0.1 to 0.05 during the training. The choice of these hyper-parameters was based on pure experimentation with different values and optimizing for the metrics described below. The Random Forest technique (Tin Kam Ho, 1998; Breiman, 2001) is a combination of decision tree predictors. Indeed, it is an approach to average numerous decision trees to obtain minimal variance. In this work, we used the Random Forest regressor with 100 decision trees and 15 maximum tree depth. The Random Forest technique is not as sensitive to the specified hyper-parameters as Neural Network approach. Another very effective technique based on decision trees is Extreme Gradient Boosting - XGBoost (Chen & Guestrin, 2016). It is an ensemble method that is developed to prevent overfitting, handle missing values, allow parallel processing, and perform cross-validation at each iteration. It tries to find an optimal output using the gradient descent algorithm to minimize the loss for the newly created model. In this work, we used XGBoost regression with squared loss, 0.05 learning rate, and 15 maximum tree depth. All the methods mentioned above are implemented and available in the libraries of the Python programming language (Van Rossum & Drake, 2009) i.e. scikit-learn (Pedregosa et al., 2011) and Keras (Chollet, 2015). Here, we have provided only a brief description. The specific set-up of the machine learning techniques used and their hyper-parameters can be found in the Jupyter notebook that is available as online material to this article (SPACE::LAB, 2020).

In order to characterize the performance of the techniques used, the following metrics were considered. The mean absolute error (MAE) represents the difference between the true label value y_i of the airglow intensity and the corresponding modeled value $\hat{y}_i$ of the i-th sample. It is defined as:

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid, \tag{1}$$

for n number of samples. Due to the reason that the absolute intensities of green and red airglow lines are different, we introduced also a relative metric the mean absolute percentage error (MAPE). It allows us to compare the performance of the techniques used for both airglow lines. Since the measured airglow intensity y_i will be always higher than zero, the MAPE is defined as:

$$MAPE(y, \hat{y}) = \frac{100\%}{n} \sum_{i=1}^{n} \frac{| y_i - \hat{y}_i |}{y_i}. \tag{2}$$

Due to the complexity of the upper atmosphere environment, the commonly used models applied for calculation of airglow intensities are limited and do not contain all the relevant processes. One of the most used, the Global Airglow (GLOW) model (Solomon et al., 1988; Solomon, 2017; Hirsch & Solomon, 2019) provides emission rates for most prominent airglow lines for particular altitude, latitude, longitude, and time. It uses energetic inputs from the Sun and aurora and also thermospheric parameters. It can also employ the output from general atmosphere circulation models such as the Thermosphere-Ionosphere-Electrodynamics General Circulation Model (TIE-GCM) (Roble et al., 1988; Qian et al., 2014). The simulated airglow brightness over the whole Earth's disk is qualitatively consistent with measurements from the most recent airglow space mission GOLD (Global-scale Observations of the Limb and Disk) (Gan et al., 2020).

3 Results and Discussion

The objective of the present work is to model the intensities of the airglow green line (557.7 nm) and red line (630.0 nm) for the period 1964–1993. For this purpose, we employed the data and techniques described in Section 2. It is noted that the main dataset was split into a subset for training and a subset for testing of each particular technique. The subsets for training and testing contain 80 % and 20 % of all data from the main dataset, respectively. The data for train and test subsets were selected randomly. The data from the main dataset are shuffled and split equally for all techniques to assure reproducible and comparable results. The comparison of the performance of the machine learning techniques used against the same subset for testing is presented in Table 2.

For the purposes of quantifying the methods' performance, the results from baseline model are also listed. They were obtained by considering simple average of the values of training labels as the modeled value $\hat{y}_i$. As expected, the lowest performance was obtained for the simplest method - linear regression. The Neural Network model provides significantly better results for MAE but even worse results for MAPE than the baseline. This is a consequence of the fact that for some low values of y_i, the modeled value of $\hat{y}_i$ might be higher by hundreds of percent although in absolute values this difference ($| y_i - \hat{y}_i |$) is not so significant. Therefore it is instructive to examine the both the MAE

Table 2. The performance of machine learning techniques used for modeling of green (557.7 nm) and red (630.0 nm) airglow lines intensities.

	I 557		**I 630**	
	MAE	**MAPE**	**MAE**	**MAPE**
Baseline	265 R	78 %	84 R	86 %
Lin. Regression	247 R	65 %	77 R	72 %
Neural Network	146 R	95 %	63 R	90 %
Random Forest	102 R	23 %	53 R	41 %
XGBoost	88 R	16 %	48 R	32 %

and MAPE metrics presented in Table 2. The evidence that the neural networks might be outperformed by techniques based on decision trees for limited datasets is well known (Wang et al., 2018). This is also the case in our work where the Random Forest technique provides lower MAE and MAPE than the Neural Network. Furthermore, the Random Forest training process was roughly $\sim$20 times computationally more efficient than the training process of the Neural Network. As the XGBoost is even more advanced than Random Forest technique, it was expected to outperform the Random Forest approach. This assumption was confirmed and the best-performing technique in our work was the XGBoost. The MAPE for green and red airglow lines were 16% and 32%, respectively. The visualization of XGBoost performance on the testing subset is displayed in Figure 2. Considering the data measurement uncertainty level of 10–15% (Fishkova, 1983), the machine learning model performs sufficiently well to qualitatively express the airglow variations. It is noted, the fact that the performance of almost all techniques is better for the green airglow line than for the red airglow line might be explained by the following consideration. The red atomic oxygen emission is strongly dependent on the electron density in the ionospheric F2 region. The green atomic oxygen emission is mainly dependent on densities of neutral species (such as O, O_2 and N_2) in the lower thermosphere. Both regions are continuously affected by various unpredictable dynamical processes determined by atmospheric waves and tidal motions. However, the amplitude of atmospheric waves and magnitude of wind velocity is higher at the altitude of the red line luminous layer. Therefore red line intensities have higher standard deviation than green line intensities. This might be reflected also in higher MAPE for red line intensities.

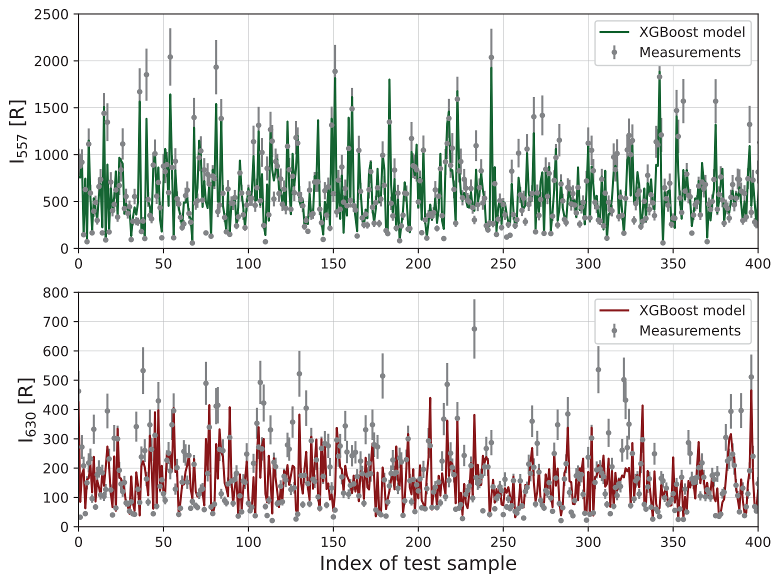


Figure 2. The performance of XGBoost model on the testing subset for green (top) and red (bottom) airglow lines intensities. The samples for the testing subset were selected randomly from all the available data. Only half of the testing subset is displayed to provide better visualization. The accuracy of the model against measurements is expressed in Table 2.

The results of the modeled intensities for green and red airglow lines over the whole studied period 1964–1993 is in Figure 3. The modeled values were obtained using all available needed input features and by the prediction of the trained machine learning model which is based on the XGBoost technique. Figure 3 represents the achievement of one of this work's goals as it contains averaged intensities of green and red airglow lines for 46,223 hours i.e. for 100% of all dark night hours within 1964–1993 period. Figure 3 serves as the visualization of the green and red airglow lines intensities variations that are displayed for a continuous period over three solar cycles. To our knowledge, airglow variation visualization for a such long period and such time resolution has not been published thus far.

To examine the credibility of the results generated by our machine learning model, we have compared them with the results of the GLOW model (Hirsch & Solomon, 2019). These results were obtained by the default setup of the GLOW model while we specified only the time, latitude and longitude. The calculated volume emission rates were integrated over all altitudes to achieve values that might be compared with the measured

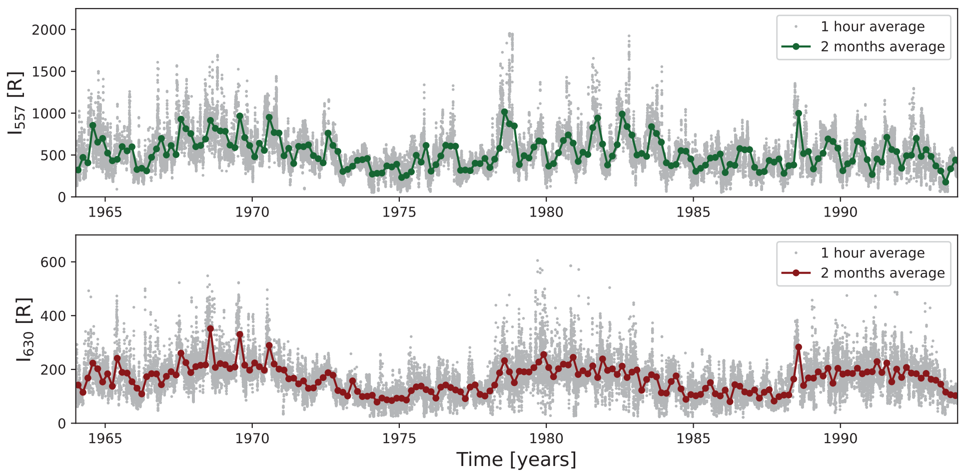


Figure 3. Visualization of airglow intensities modeled by the XGBoost technique for the location Abastumani (Georgia) over the years 1964–1993. The averages over 1 hour and 2 months for green *(top)* and red *(bottom)* airglow lines are displayed. Only dark night hours are considered.

airglow data. For the same testing dataset as was used for Table 2, the GLOW model achieved as follows for the green line: MAE equals 280 R and MAPE equals 89%, for the red line: MAE equals 109 R and MAPE equals 84%. These values are not as good as the results of our machine learning model. This can be explained by the fact that the particular measured data might be influenced by phenomena that are not considered in the default setup of the GLOW model. The performance of our machine learning model and the GLOW model is presented in Figure 4 for the full period 1964–1993. This shows that both models are qualitatively in good agreement. The correlation coefficients of simulated intensities for the GLOW model and our machine learning model based on XGBoost averaged over 2 months and considering a linear least-squares regression are 0.48 and 0.54 for green and red line, respectively. It is an important result that the data-driven model can provide valuable results even with a comparison of the physical model generally used. Even-more, as displayed in Figure 4, the data-driven model is less uniform than the physical model and might be more consistent with the real variability expressed by the measurements. However, it is important to note, the GLOW model is much more general than the particular data-driven model and can be used for any location and time because it does not require any measured airglow data for the input.

To examine the performance of our data-driven model for the completely unseen time period, we made an experiment where we split the main dataset for the subset for training and testing covering the years 1964–1979 (i.e. 50% of the previously used dataset) and for the subset for validation covering the years 1979–1993. The new model was trained and tested by using the training and testing subsets only. Its performance was then investigated by the validation subset. The MAE and MAPE for the green airglow line were 298 R and 99%, respectively. The MAE and MAPE for the red airglow line were 90 R and 95%, respectively. The mean errors are significantly higher than values in Table 2 but this was expected because we used only data from a 15-year period for the training and testing process. The metrics for the GLOW model by using the same validation dataset were very similar. The MAE and MAPE for the green airglow line were 290 R and 105%, respectively. The MAE and MAPE for the red airglow line were 119 R and 100%, respectively. This demonstrates that for a completely unseen time period our data-driven approach is still able to produce comparable results to the GLOW model. The correlation coefficients are now equal to 0.46 and 0.8 for the green and red line, respectively. It is interesting that the correlation coefficient for the red line is now significantly higher. This means that when we used less data for training of our model its results for the red line have a greater similarity to the results of the GLOW model. It is rather a surprising result, because it might be expected that for the smaller training dataset the data-driven model would depart more from the GLOW model. The obvious explanation is that the GLOW model as well as our model trained on only a 15-year period do not consider all the phenomena that influence atomic oxygen airglow emissions. There is also a possible explanation that airglow measurements acquired during the solar cycle number 22 (1986–1996) were somehow different from the data acquired during the previous two solar cycles. This can be caused by the unknown contamination of the data or by occurrence of some unique processes that produced unexpected airglow intensities. We will investigate this inconsistency in the future by comparison with airglow measurements from other locations for the similar time period.

Another examination of the credibility of our machine learning model is its ability to express the airglow variations briefly presented in Section 1. As all inputs for the data-driven model are modulated directly or indirectly by the cycle of solar activity and the seasons, it is not a surprise that these variations should be present also in modeled airglow intensities. It is examined if the characteristics of these variations are compat-

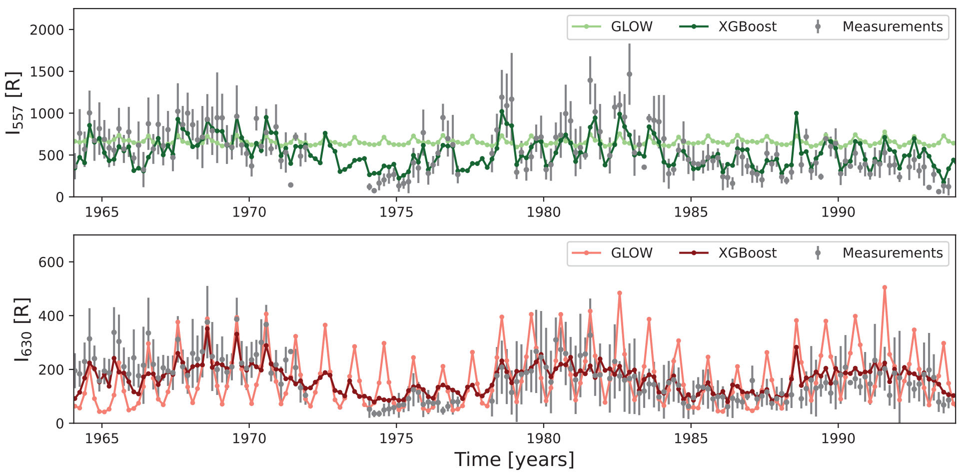


Figure 4. The time series of green *(top)* and red *(bottom)* airglow lines for the period 1964–1993. The 2-month averages of calculated intensities using the GLOW model and our data-driven model based on the XGBoost technique. The 2-month averages of measurements from Abastumani (Georgia) (see Figure 1) are displayed together with their standard deviations.

ible with the results of other authors. The airglow modulation by an 11-year solar cycle is visible in Figure 3 at a glance. The green and red airglow lines intensities are maximal for the periods around the maxima of solar activity in the years 1969, 1980, and 1991, which is consistent with results presented in studies (e.g. Deutsch & Hernandez, 2003; Reid et al., 2014). The annual variation can be also recognized in Figure 3. According to previous studies (Shepherd et al., 2006) this variation of green line intensities should have its minimum in spring and maximum in autumn for the considered location in the middle latitudes of the northern hemisphere. The results of our data-driven model presented in Figure 5 *(*top) are consistent with these studies. The assumption for the red airglow line for the considered location is that the maximum average intensity should be in summer and the minimum near equinoxes (Khomich et al., 2008). The results presented in Figure 5 *(bottom)* are also consistent with this assumption. We note, there are many more airglow variations present in Figures 3 and 5. They might be recognized by further investigation of the developed data-driven model results. These analyses and comparison with various measurements, as done by other authors (e.g. Deutsch & Hernandez, 2003; Gudadze et al., 2008), are objectives for future publication.

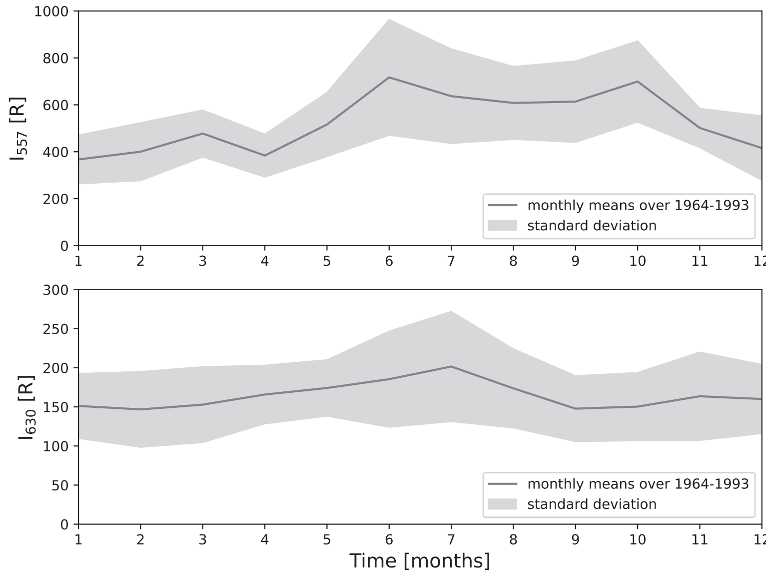


Figure 5. The average intensities calculated by a data-driven model based on the XGBoost technique for Abastumani (Georgia). The intensities were averaged over a particular month and for the years 1964–1993. The standard deviations from the mean values over the years are also displayed.

4 Conclusions

Space data are of irreplaceable value as they provide information about phenomena that can not be repeated. However, the occurrence of missing measurements and gaps in the time series is very common. This is especially true for the ground-based measurements where the observations are limited by the weather conditions. We have used the most recent machine learning techniques to solve the regression problem and to model the missing intensities of green and red airglow lines for the location Abastumani (Georgia) over the time period 1964–1993. For this purpose, a data-driven approach was used. The photometric airglow measurements were used as the labels (target outputs) and space weather indices, thermosphere-ionosphere parameters, and Sun-Earth distances were used as the features (inputs). The techniques of Linear Regression, Neural Network, Random Forest, and XGBoost were employed and their performance was compared against the testing dataset. The model based on the XGBoost technique outperformed the others and provided mean absolute percentage error (MAPE) of 16% and 32% for the green and red airglow lines, respectively. This performance is sufficient to qualitatively express the overall airglow variation, and enables the modeled data to represent the missing measurements with a reasonable level of uncertainty. The obtained data visualize the variations in the intensities of the green and red airglow lines over the period of three solar cycles. The results from the data-driven model are consistent with the GLOW model (Solomon, 2017) and depict the main variations related to solar activity and the seasons.

The modeled airglow data might contribute to understanding the processes in the interface region between the space environment and Earth's atmosphere. Even more, the absolute values of airglow intensities and the range of their variation are crucial for future missions like EUSO-SPB2 (Wiencke, 2019) and POEMMA (NASA Probe Study report, 2020; Anchordoqui et al., 2020). These missions are designed to observe extensive air showers induced by ultra-high energy cosmic rays and to observe Cherenkov light induced by cosmic neutrinos. Indeed, airglow emissions set the energy threshold of the events that could be recognized in the Earth's night atmosphere by observation from orbit (JEM-EUSO collaboration, 2019; Krizmanic, 2021). For this purpose we plan to extend the visualization of the airglow intensities for the years 1994–2020 as the input features should be available. We would like to also focus on the short time periods when the airglow intensities were significantly high and to investigate the possible explanations of these specific events.

Acknowledgments

These airglow studies are supported by the government of Slovakia through the ESA contracts No. 4000125330/18/NL/SC and No. 4000125987/18/NL/SC under the PECS (Plan for European Cooperating States). ESA Disclaimer: The view expressed herein can in no way be taken to reflect the official opinion of the European Space Agency. The studies of Sun-Earth connection are supported by the VEGA grant agency project 2/0155/18. The work is also supported by the Slovak Academy of Sciences MVTS JEM-EUSO grant. The studies related to machine learning techniques are supported by GlobalLogic Slovakia s.r.o. G. Didebulidze is supported by the Georgian Shota Rustaveli National Science Foundation Grant no. FR17–357. Data Availability Statement: The airglow data were provided by G. Didebulidze and are publicly available through the NDMC database (https://ndmc.dlr.de). The data of space weather, thermosphere, and ionosphere parameters are publicly available through the NASA data centers (https://omniweb.gsfc.nasa.gov,

https://ccmc.gsfc.nasa.gov). The presented results can be reproduced by the Jupyter notebook publicly available at https://doi.org/10.5281/zenodo.4306913.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* (Software available from tensorflow.org)

Anchordoqui, L. A., Bergman, D. R., Bertaina, M. E., Fenu, F., Krizmanic, J. F., Liberatore, A., … Wiencke, L. (2020, January). Performance and science reach of the Probe of Extreme Multimessenger Astrophysics for ultrahigh-energy particles. *Physical Review D*, *101*(2), 023012. doi: 10.1103/PhysRevD.101.023012

Bag, T., Singh, V., & Krishna, M. S. (2017). Study of atomic oxygen greenline dayglow emission in thermosphere during geomagnetic storm conditions. *Advances in Space Research*, *59*(1), 302 - 310. doi: https://doi.org/10.1016/j.asr.2016.08.037

Ball, N. M., & Brunner, R. J. (2010, January). Data Mining and Machine Learning in Astronomy. *International Journal of Modern Physics D*, *19*(7), 1049-1106. doi: 10.1142/S0218271810017160

Bilitza, D., Altadill, D., Truhlik, V., Shubin, V., Galkin, I., Reinisch, B., & Huang, X. (2017). International reference ionosphere 2016: From ionospheric climate to real-time weather predictions. *Space Weather*, *15*(2), 418-429. doi: https://doi.org/10.1002/2016SW001593

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. doi: 10.1023/A:1010933404324

Camporeale, E., Wing, S., & Johnson, J. (2018). *Machine learning techniques for space weather*. Elsevier. doi: https://doi.org/10.1016/C2016-0-01976-9

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). New York, NY, USA: ACM. doi: 10.1145/2939672.2939785

Chollet, F. e. a. (2015). *Keras.* `https://keras.io`.

Cybenko, G. (1989, December 1). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, *2*(4), 303–314. doi: 10.1007/BF02551274

Deutsch, K. A., & Hernandez, G. (2003, December). Long-term behavior of the OI 558 nm emission in the night sky and its aeronomical implications. *Journal of Geophysical Research (Space Physics)*, *108*, 1430. doi: 10.1029/2002JA009611

Didebulidze, G. G., Lomidze, L. N., Gudadze, N. B., Pataraya, A. D., & Todua, M. (2011). Long-term changes in the nightly behaviour of the oxygen red 630.0 nm line nightglow intensity and trends in the thermospheric meridional wind velocity. *International Journal of Remote Sensing*, *32*(11), 3093-3114. doi: 10.1080/01431161.2010.541523

Eastes, R. W., McClintock, W. E., Burns, A. G., Anderson, D. N., Andersson, L., Codrescu, M., … Oberheide, J. (2017, October). The Global-Scale Observations of the Limb and Disk (GOLD) Mission. *Space Sci. Rev.*, *212*, 383-408. doi: 10.1007/s11214-017-0392-2

Eastes, R. W., Solomon, S. C., Daniell, R. E., Anderson, D. N., Burns, A. G., England, S. L., … McClintock, W. E. (2019). Global-scale observations of the equatorial ionization anomaly. *Geophysical Research Letters*, *46*(16), 9318-9326. doi: 10.1029/2019GL084199

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*(15), 3133-3181.

Fishkova, L. M. (1983). *The night airglow of the earth mid-latitude upper atmosphere.*

Gan, Q., Eastes, R. W., Burns, A. G., Wang, W., Qian, L., Solomon, S. C., … McClintock, W. E. (2020). First synoptic observations of geomagnetic storm effects on the global-scale oi 135.6-nm dayglow in the thermosphere by the gold mission. *Geophysical Research Letters*, *47*(3), e2019GL085400. doi: 10.1029/2019GL085400

George, D., & Huerta, E. (2018). Deep learning for real-time gravitational wave detection and parameter estimation: Results with advanced ligo data. *Physics Letters B*, *778*, 64 - 70. doi: https://doi.org/10.1016/j.physletb.2017.12.053

Gudadze, N. B., Didebulidze, G. G., Javakhishvili, G. S., Shepherd, M. G., & Vardosanidze, M. V. (2007, February). Long-term variations of the oxygen red 630 nm line nightglow intensity. *Canadian Journal of Physics*, *85*(2), 189-198. doi: 10.1139/P07-032

Gudadze, N. B., Didebulidze, G. G., Lomidze, L. N., Javakhishvili, G. S., Marsagishvili, M. A., & Todua, M. (2008). Different long-term trends of the oxygen red 630.0 nm line nightglow intensity as the result of lowering the ionosphere f2 layer. *Annales Geophysicae*, *26*(8), 2069–2080. doi: 10.5194/angeo-26-2069-2008

Hannawald, P., Schmidt, C., Sedlak, R., Wüst, S., & Bittner, M. (2019). Seasonal and intra-diurnal variability of small-scale gravity waves in oh airglow at two alpine stations. *Atmospheric Measurement Techniques*, *12*(1), 457–469. doi: 10.5194/amt-12-457-2019

Hirsch, M., & Solomon, S. (2019, September). *space-physics/ncar-glow.* Zenodo. doi: 10.5281/zenodo.3463662

Immel, T. J., England, S. L., Mende, S. B., Heelis, R. A., Englert, C. R., Edelstein, J., ... Sirk, M. M. (2018, February). The Ionospheric Connection Explorer Mission: Mission Goals and Design. *Space Sci. Rev.*, *214*, 13. doi: 10.1007/s11214-017-0449-2

Isobe, T., Feigelson, E. D., Akritas, M. G., & Babu, G. J. (1990, November). Linear Regression in Astronomy. I. *Astrophys. J.*, *364*, 104. doi: 10.1086/169390

JEM-EUSO collaboration. (2019, September). Ultra-violet imaging of the night-time earth by EUSO-Balloon towards space-based ultra-high energy cosmic ray observations. *Astroparticle Physics*, *111*, 54-71. doi: 10.1016/j.astropartphys.2018.10.008

Khomich, V. Y., Semenov, A. I., & Shefov, N. N. (2008). *Airglow as an Indicator of Upper Atmospheric Structure and Dynamics.* Springer-Verlag.

King, J. H., & Papitashvili, N. E. (2005). Solar wind spatial scales in and comparisons of hourly wind and ace plasma and magnetic field data. *Journal of Geophysical Research: Space Physics*, *110*(A2). doi: 10.1029/2004JA010649

Krizmanic, J. F. (2021). Space-based extensive air shower optical cherenkov and fluorescence measurements using sipm detectors in context of poemma. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *985*, 164614. doi: https://doi.org/10.1016/j.nima.2020.164614

Leonovich, L. A., Mikhalev, A. V., & Leonovich, V. A. (2011, Aug 17). The 557.7 and 630-nm atomic oxygen midlatitude airglow variations associated with geomagnetic activity. *Atmospheric and Oceanic Optics*, *24*(4), 396. doi: 10.1134/S1024856011040105

Liu, G., & Shepherd, G. G. (2008). An investigation of the solar cycle impact on the lower thermosphere o(1s) nightglow emission as observed by windii/uars. *Advances in Space Research*, *42*(5), 933 - 938. doi: https://doi.org/10.1016/j.asr.2007.10.008

Liu, G., Shepherd, G. G., & Roble, R. G. (2008). Seasonal variations of the nighttime o(1s) and oh airglow emission rates at mid-to-high latitudes in the context of the large-scale circulation. *Journal of Geophysical Research: Space Physics*, *113*(A6). doi: 10.1029/2007JA012854

Mackovjak, Š., Bobík, P., Baláž, J., Strhárský, I., Putiš, M., & Gorodetzky, P. (2019, April). Airglow monitoring by one-pixel detector. *Nuclear Instruments and Methods in Physics Research A*, *922*, 150-156. doi: 10.1016/j.nima.2018.12.073

Makela, J. J., Harding, B. J., Meriwether, J. W., Mesquita, R., Sanders, S., Ridley, A. J., … Martinis, C. R. (2014). Storm time response of the midlatitude thermosphere: Observations from a network of fabry-perot interferometers. *Journal of Geophysical Research: Space Physics*, *119*(8), 6758-6773. doi: 10.1002/2014JA019832

NASA Probe Study report. (2020). *POEMMA: Probe of Extreme Multi-Messenger Astrophysics .* `https://smd-prod.s3.amazonaws.com/science-pink/s3fs-public/atoms/files/1_POEMMA_Study_Rpt_0.pdf`.

NDMC. ((last access: November 30, 2020)). *The Network for the Detection of Mesospheric Change (NDMC), available at.* `https://ndmc.dlr.de`.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., … Lerer, A. (2017). *Automatic differentiation in pytorch.*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pfaff, R. F. (2012, June). The Near-Earth Plasma Environment. *Space Sci. Rev.*, *168*, 23-112. doi: 10.1007/s11214-012-9872-6

Picone, J. M., Hedin, A. E., Drob, D. P., & Aikin, A. C. (2002). Nrlmsise-00 empirical model of the atmosphere: Statistical comparisons and scientific issues. *Journal of Geophysical Research: Space Physics*, *107*(A12), SIA 15-1-SIA 15-16. doi: 10.1029/2002JA009430

Qian, L., Burns, A. G., Emery, B. A., Foster, B., Lu, G., Maute, A., ... Wang, W. (2014). The ncar tie-gcm. In *Modeling the ionosphere–thermosphere system* (p. 73-83). American Geophysical Union (AGU). doi: 10.1002/9781118704417.ch7

Rayleigh, L., & Jones, H. S. (1935, Aug). The Light of the Night-Sky: Analysis of the Intensity Variations at Three Stations. *Proceedings of the Royal Society of London Series A*, *151*(872), 22-55. doi: 10.1098/rspa.1935.0133

Reid, I. M., Spargo, A. J., & Woithe, J. M. (2014). Seasonal variations of the nighttime O (1S) and OH (8-3) airglow intensity at Adelaide, Australia. *Journal of Geophysical Research: Atmospheres*, *119*(11), 6991–7013. doi: 10.1002/2013JD020906

Roble, R. G., Ridley, E. C., Richmond, A. D., & Dickinson, R. E. (1988). A coupled thermosphere/ionosphere general circulation model. *Geophysical Research Letters*, *15*(12), 1325-1328. doi: 10.1029/GL015i012p01325

Shepherd, G., Thuillier, G., Gault, W., Solheim, B., Hersom, C., Alunni, J., ... Wimperis, J. (1993, 06). Windii, the wind imaging interferometer on the upper atmosphere research satellite. *J. Geophys. Res.*, *98*. doi: 10.1029/93JD00227

Shepherd, G. G., Cho, Y.-M., Liu, G., Shepherd, M. G., & Roble, R. G. (2006, December). Airglow variability in the context of the global mesospheric circulation. *Journal of Atmospheric and Solar-Terrestrial Physics*, *68*, 2000-2011. doi: 10.1016/j.jastp.2006.06.006

Silverman, S. M. (1970, October). Night Airglow Phenomenology. *Space Sci. Rev.*, *11*, 341-379. doi: 10.1007/BF00241526

Solomon, S. C. (2017). Global modeling of thermospheric airglow in the far ultraviolet. *Journal of Geophysical Research: Space Physics*, *122*(7), 7834-7848. doi: 10.1002/2017JA024314

Solomon, S. C., Hays, P. B., & Abreu, V. J. (1988). The auroral 6300 Å emission: Observations and modeling. *Journal of Geophysical Research: Space Physics*, *93*(A9), 9867-9882. doi: 10.1029/JA093iA09p09867

SPACE::LAB. (2020, December). *space-lab-sk/airglow_data-driven_model: First release.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.4306913 doi: 10.5281/zenodo.4306913

Tin Kam Ho. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832-844.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. Scotts Valley, CA: CreateSpace.

von Savigny, C. (2017, Oct 17). Airglow in the earth atmosphere: basic characteristics and excitation mechanisms. *ChemTexts*, *3*(4), 14. doi: 10.1007/s40828-017-0051-y

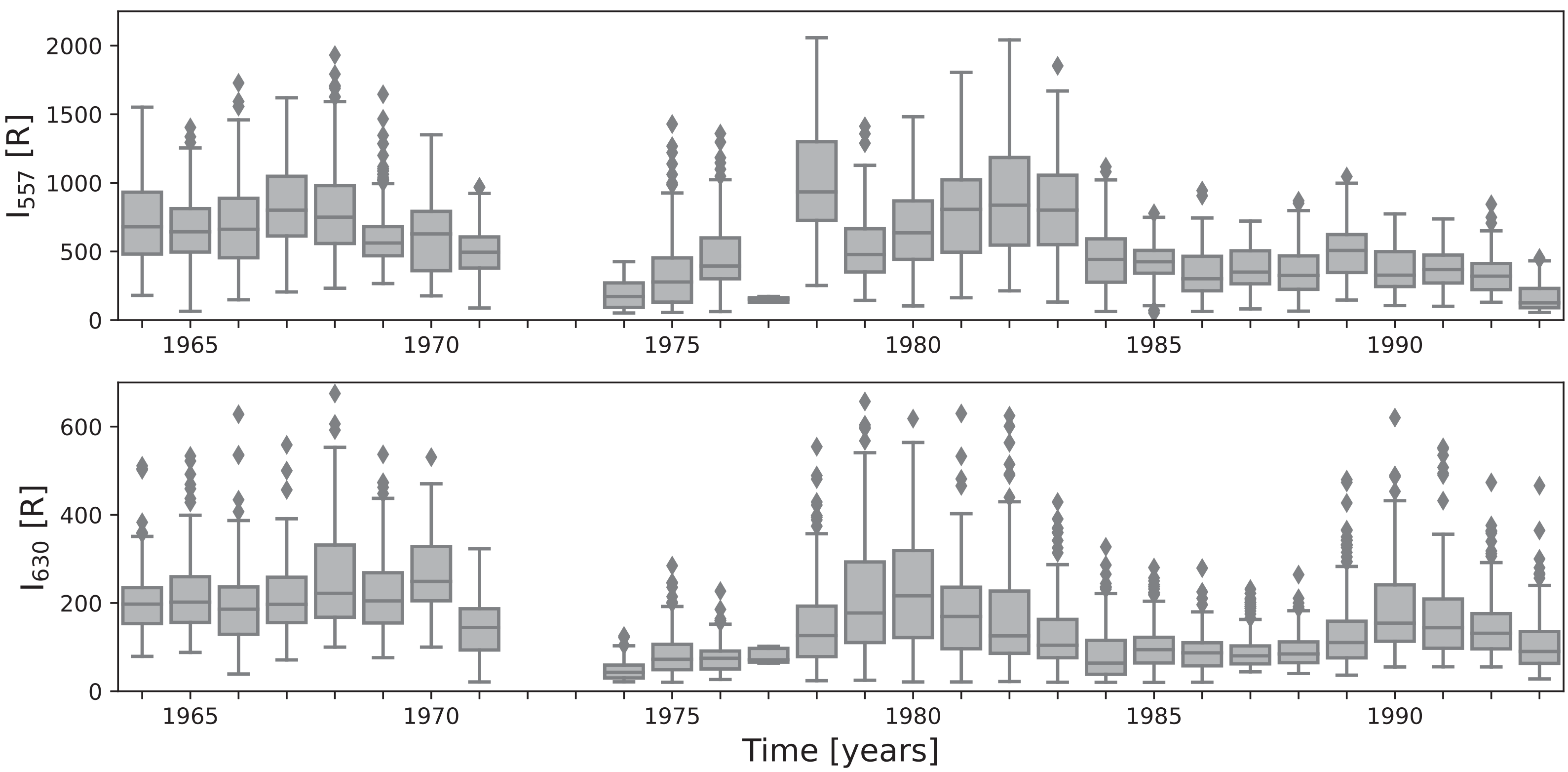
Wang, S., Aggarwal, C., & Liu, H. (2018, 10). Random-forest inspired neural networks. *ACM Transactions on Intelligent Systems and Technology*, *9*. doi: 10.1145/3232230

Wiencke, L. (2019, July). The Extreme Universe Space Observatory on a Super-Pressure Balloon II Mission. In *36th international cosmic ray conference (icrc2019)* (Vol. 36, p. 466).

Wüst, S., Schmidt, C., Hannawald, P., Bittner, M., Mlynczak, M. G., & Russell III, J. M. (2019). Observations of oh airglow from ground, aircraft, and satellite: investigation of wave-like structures before a minor stratospheric warming. *Atmospheric Chemistry and Physics*, *19*(9), 6401–6418. doi: 10.5194/acp-19-6401-2019

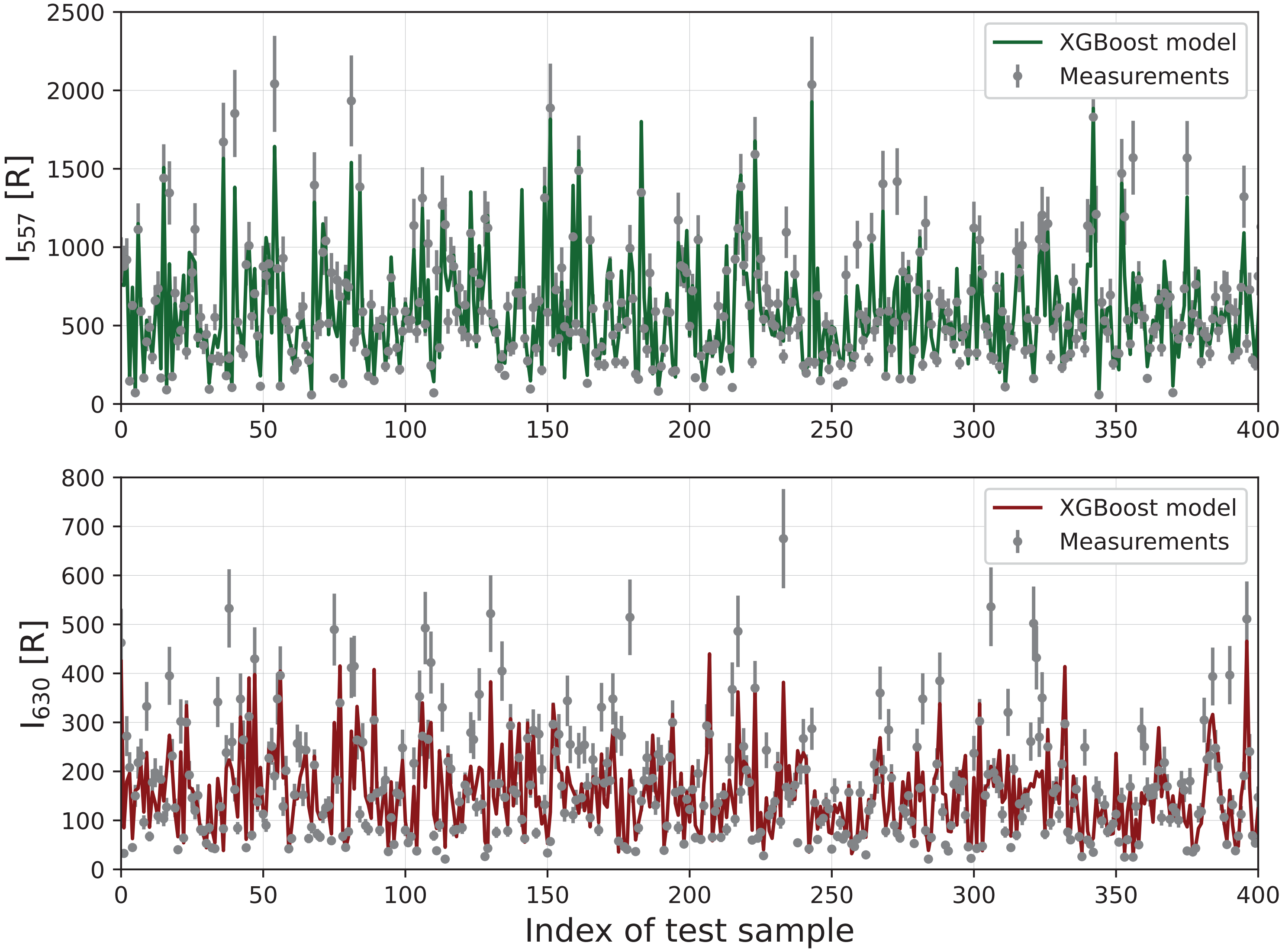
Zucker, S., & Giryes, R. (2018, mar). Shallow transits—deep learning. i. feasibility study of deep learning to detect periodic transits of exoplanets. *The Astronomical Journal*, *155*(4), 147. doi: 10.3847/1538-3881/aaae05

figure1.eps.



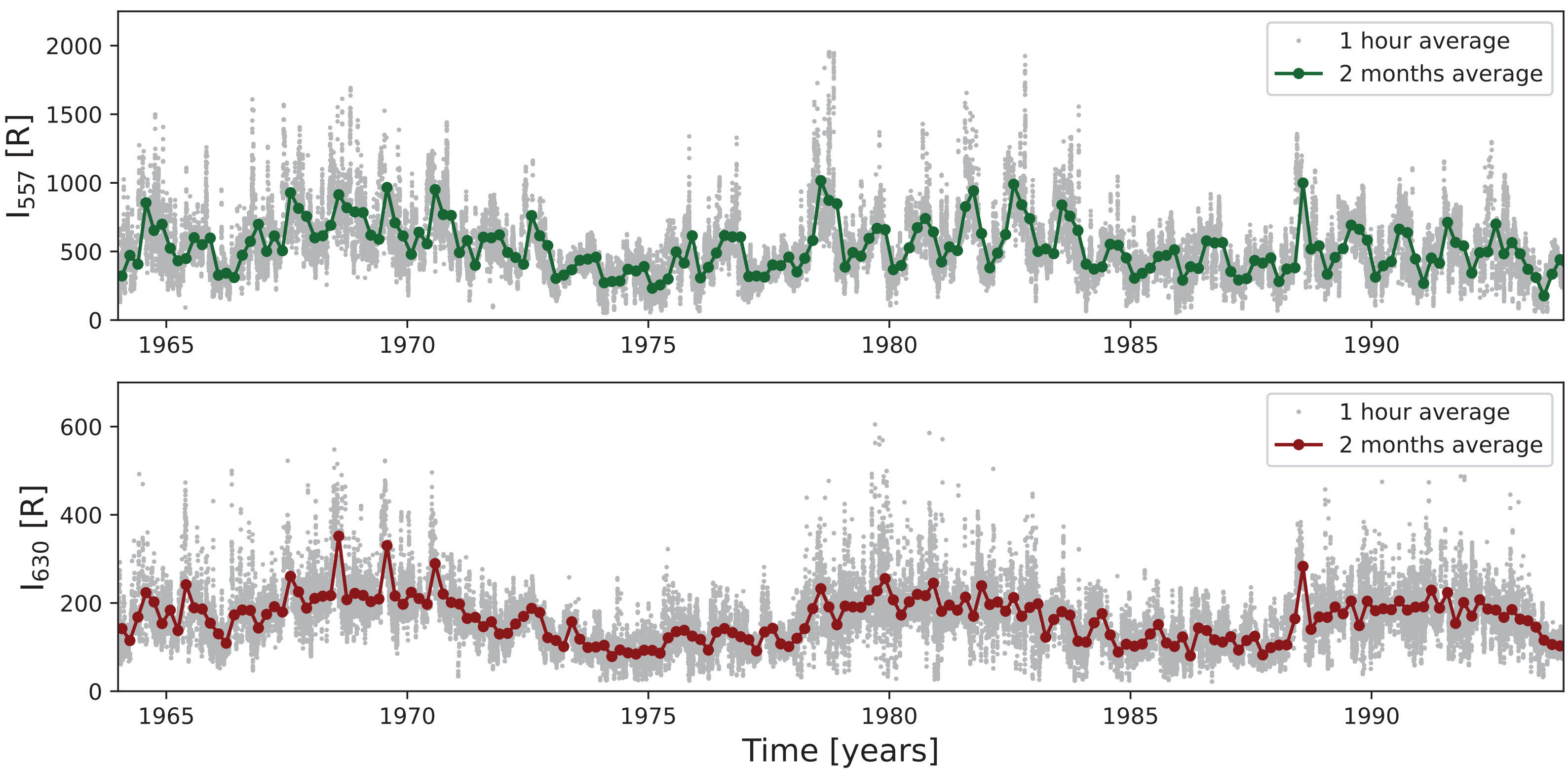
2000
1500
1000
500
0
I_{557} [R]
1965
1970
1975
1980
1985
1990
600
400
200
0
I_{630} [R]
1965
1970
1975
1980
1985
1990
Time [years]

figure2.eps.



2500
2000
1500
1000
500
0
I_{557} [R]
XGBoost model
Measurements
0
50
100
150
200
250
300
350
400
800
700
600
500
400
300
200
100
0
I_{630} [R]
XGBoost model
Measurements
0
50
100
150
200
250
300
350
400
Index of test sample

figure3.eps.



1 hour average
2 months average
I_{557} [R]
2000
1500
1000
500
0
1965
1970
1975
1980
1985
1990
1 hour average
2 months average
I_{630} [R]
600
400
200
0
1965
1970
1975
1980
1985
1990
Time [years]

figure4.eps.

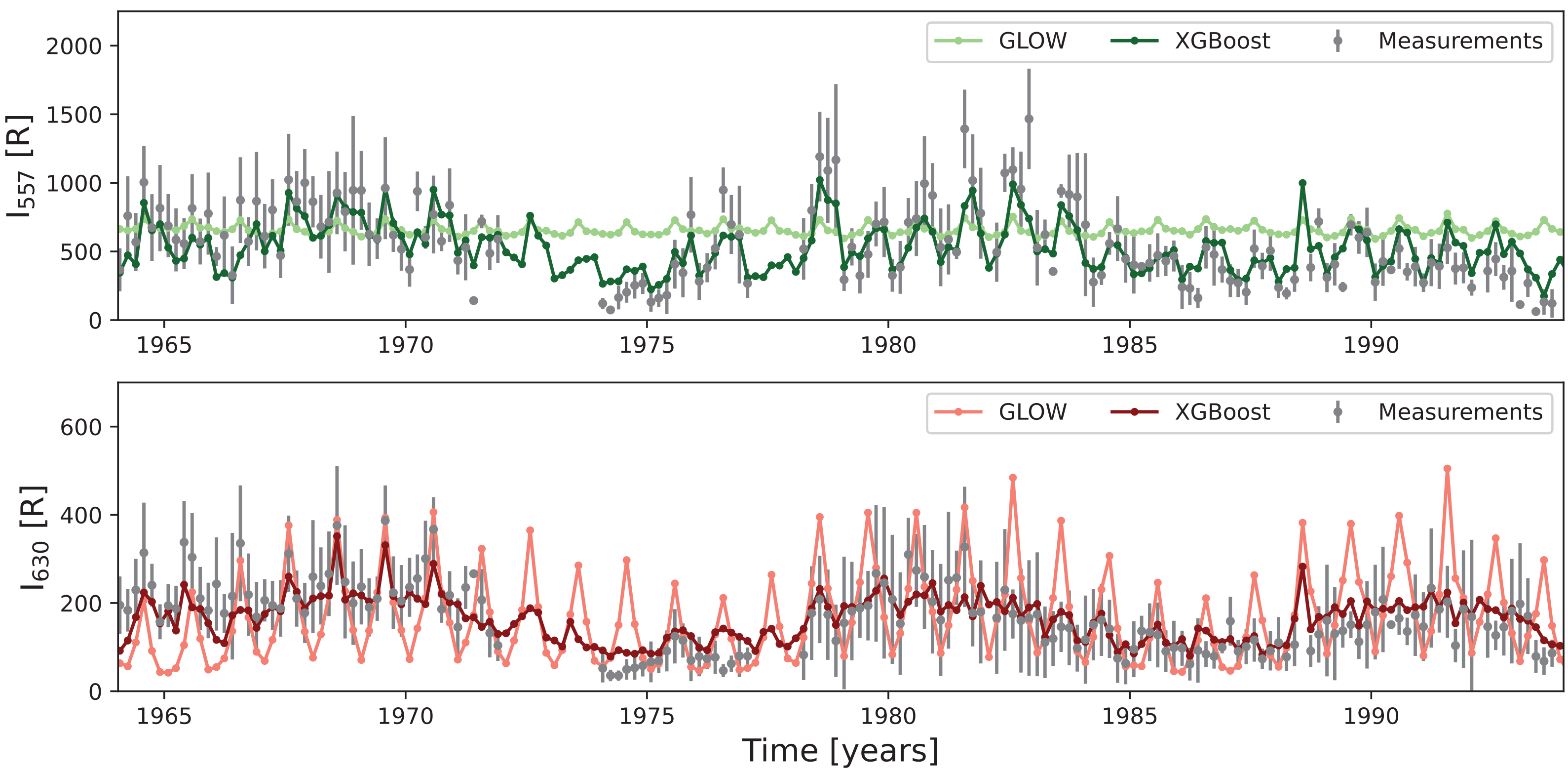
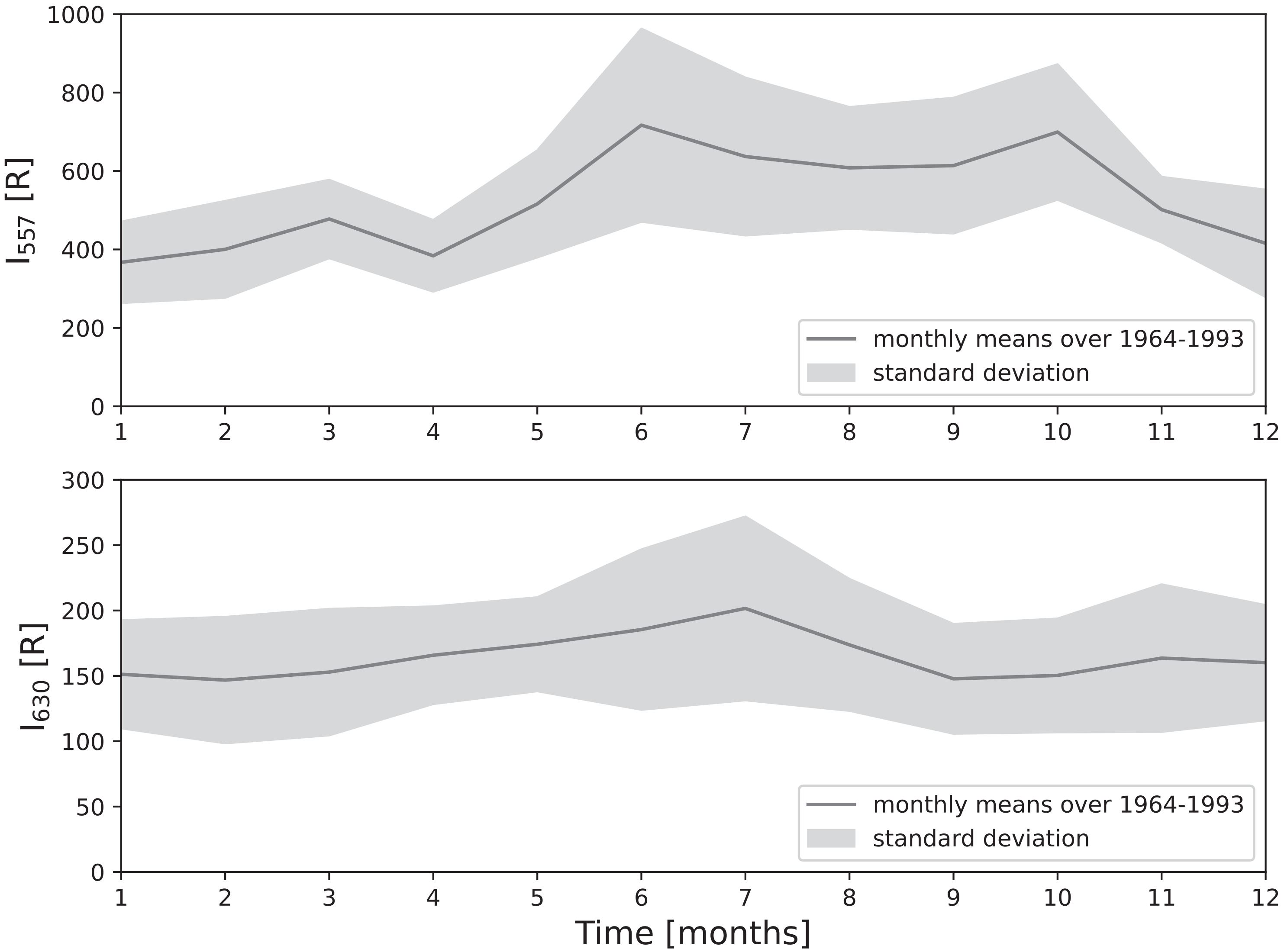

GLOW
XGBoost
Measurements
2000
1500
1000
500
0
I557 [R]
1965
1970
1975
1980
1985
1990
600
400
200
I630 [R]
Time [years]

figure5.eps.


1000
800
600
400
200
0
I_{557} [R]
monthly means over 1964-1993
standard deviation
1
2
3
4
5
6
7
8
9
10
11
12
300
250
200
150
100
50
0
I_{630} [R]
monthly means over 1964-1993
standard deviation
Time [months]