Generalized Additive Models for Solar Irradiance Forecasting

Daniel A. Brandt [1] Erick F. Vega [1] Aaron J. Ridley [2]

[1]Michigan Tech Research Institute [2]University of Michigan

Introduction

Solar irradiance forecasting is of great importance for the space weather community, particularly for applications in reentry prediction, satellite collision avoidance, and characterizing solar-terrestrial climatology. Proxies for solar irradiance are routinely used in atmospheric modeling, though recent modeling developments have begun incorporating the outputs of empirical models of solar spectra in numerous wavelength bins, such as the Flare Irradiance Spectral Model 2, which models solar spectral irradiance between 0.1-190 nm [1]. Improvements to these modeling paradigms may be achieved through the leveraging of well-grounded statistical techniques, particularly those of a Bayesian character. To this end, we present a preliminary application of Generalized Additive Models (GAMs) [2] used in conjunction with multiple linear regression to model solar spectral irradiance as a function of Day-of-the-Year and the solar indices Lyman-α, F10.7, and sunspot number (R). The GAM is trained on Solar Cycle 22 and the ascending phase of Solar Cycle 23 using FISM2 data, and multiple linear regression on residuals is performed on the descending phase of Solar Cycle 23 using TIMED/SEE data. The results are presented for solar irradiance at 425 Å for the ascending phase of Solar Cycle 24, and are generalizable to any wavelength.

Methodology

Solar irradiance data outputs from FISM2 and measurements from TIMED/SEE are obtained from LISIRD, while solar indices are obtained from NASA's OMNIWeb Data Explorer. These data are processed before the fitting of a GAM and the application of multiple linear regression.

- **Resampling:** FISM2 daily-averaged solar spectral irradiance is resampled to the hourly resolution of NASA OMNIWeb solar indices Lyman-α, F10.7, and R.
- **Data Segmentation:** FISM2 and solar index data is segmented into three temporal regions: (1) A training region for initial model fitting, (2) a training region for modeling residuals, and (3) a test region for assessing performance of the final model.
- **Initial Model Fitting:** In the first training region, a GAM is fit between solar indices and FISM2 irradiances. This model we term $Y_{\text{initial}}(\bar{x})$, where $\bar{x}$ is a vector of solar indices.
- **Residuals Regression Analysis:** In the second training region, multiple linear regression is performed between the differences between outputs of $Y_{\text{initial}}(\bar{x})$ and SEE irradiances, and solar indices, generating a model for residuals $\delta(\bar{x})$.
- **Final Model Prediction:** In the test region, the final model $Y_{\text{final}}(\bar{x}) = Y_{\text{initial}}(\bar{x}) + \delta(\bar{x})$ is driven by historical solar index data to model the ascending phase of Solar Cycle 24.

Statistical Approach in Detail

Generalized linear models (GAMs) are a mature statistical approach for modeling complex non-linear relationships. The particular functional form used in this work is contrasted with conventional linear models below:

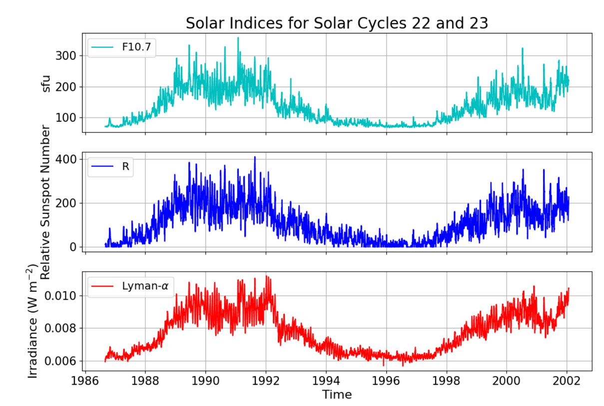
$$\underbrace{Y=\sum \beta_i X_i}_{\text{conventional linear model}} \quad \text{vs.} \quad \underbrace{Y=\sum \overbrace{h_i(X_i)}^{\text{1D spline}}}_{\text{generalized additive model}}$$

The response variable Y, i.e., the irradiance spectra at fixed wavelengths, is modeled as a sum of (non-linear) spline functions that are independently regularized. Note that higher-order multiplicative interactions between the spectra and the predictors can also be captured within this framework with the inclusion of cross terms $h_{ij}(X_i X_j)$.

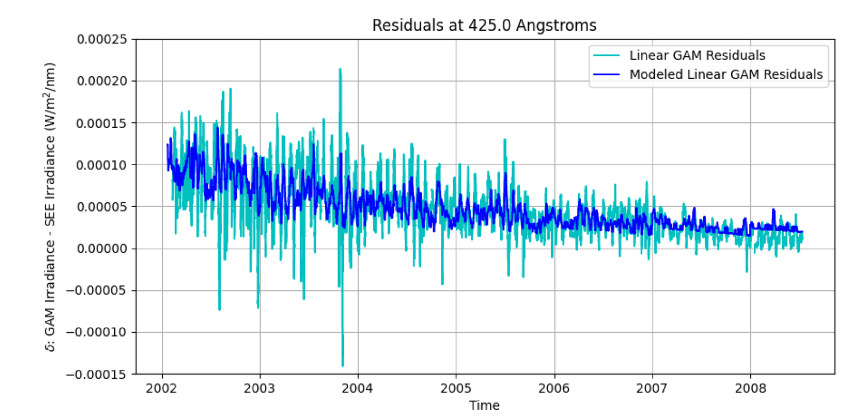
The advantages of the proposed approach are (1) **flexibility**, a wide class of functions can be approximated using GAMs, thereby enabling the learning of complex dependencies, (2) **explainability**, by analyzing the marginal functions $h_i(X_i)$, contributions to the response from each of the covariates can be precisely quantified and (3) **generalizability**, the parametric nature of the model make it amenable to principled error analysis and confidence bound estimation.

Initial Model Fitting

Solar indices Lyman-α, F10.7, and sunspot number (R) were specifically selected by reason of their serving as reasonable proxies for the solar chromosphere, corona, and photosphere, respectively.



(a) Solar indices.



(b) Residuals between Y_{initial} and FISM2 during the descending phase of Solar Cycle 23 (cyan), and the modeled residuals found via $\delta(\bar{x})$ (blue).

Figure 1. Solar indices for Solar Cycle 22 and the ascending phase of Solar Cycle 23 (left), and residuals between the initial model fit and FISM2 data (right).

A GAM (Y_{initial}) linear in spline terms for each solar index was fit for the FISM2 data. Residuals of Y_{initial} were regressed against solar indices $\bar{x}$ during the declining phase of Solar Cycle 23. Multiple linear regression was then performed between Y_{initial} residuals with respect to FISM2 data and $\bar{x}$ during the declining phase of Solar Cycle 23. A linear model was constructed for the residuals, due to their fairly stationary behavior (Figure 1b). For the wavelength considered, the final model $Y_{\text{final}}(\bar{x})$ is then found as $Y_{\text{final}}(\bar{x}) = Y_{\text{initial}}(\bar{x}) + \delta(\bar{x})$.

Results

Predictions for the Ascending Phase of Solar Cycle 24

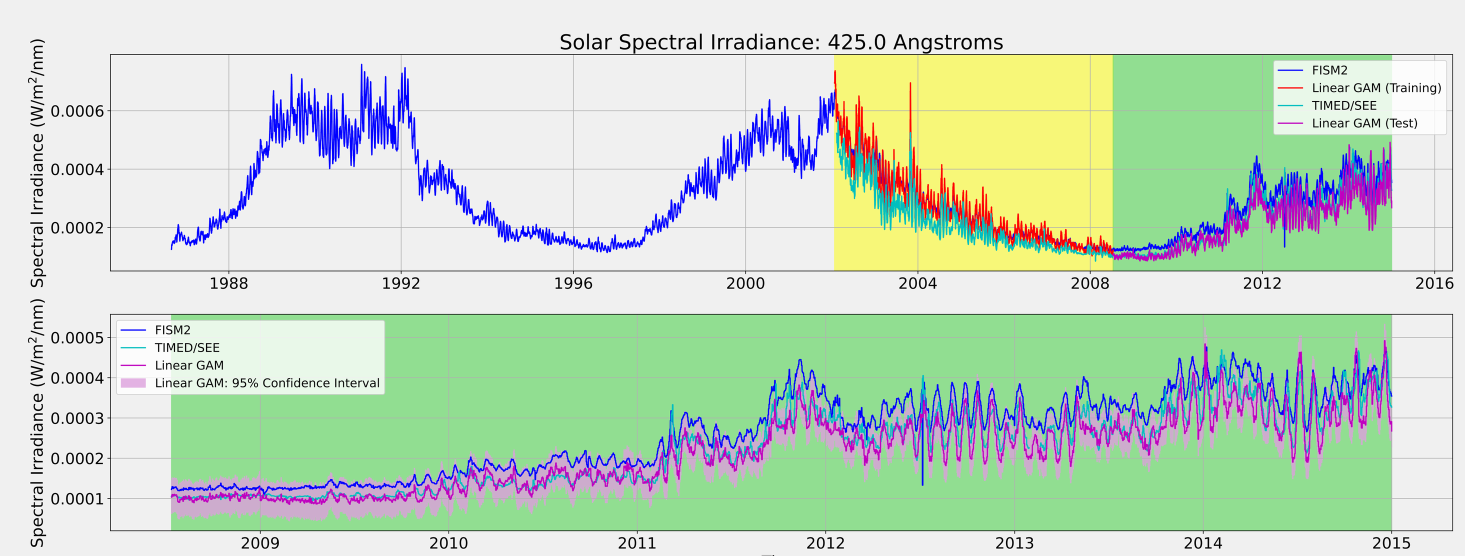
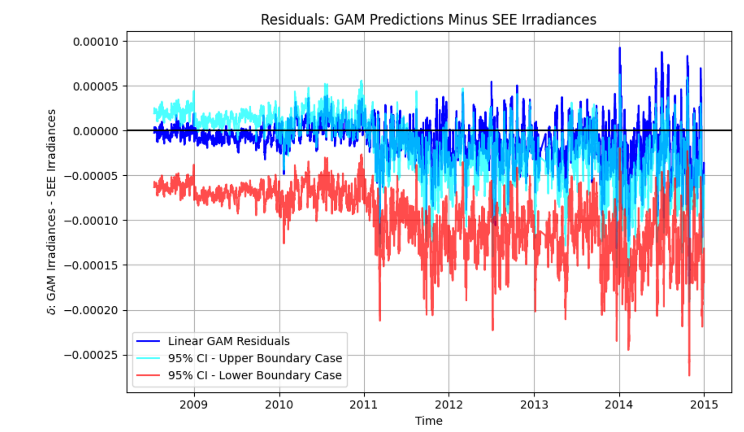


Figure 2. GAM results for the training period (yellow region) and the test period (green). The leftmost unshaded white region comprises of FISM2 data only, which was used to fit Y_{initial}. The yellow region corresponds to where $\delta(\bar{x})$ was determined, and the green region where $Y_{\text{final}}(\bar{x})$ was applied.

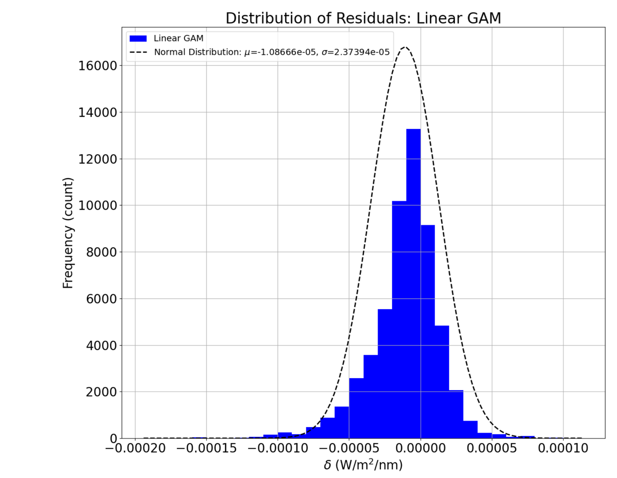
Despite the initial fitting of Y_{initial} on only 1.5 solar cycles of FISM2 data, and the modeling of residuals on only the declining phase of a single solar cycle, the performance of the final model $Y_{\text{final}}(\bar{x})$ during the ascending phase of Solar Cycle 24 (using historical solar indices) is respectable. Percent error between Y_{final} and TIMED/SEE was consistently lower than FISM2, with an average value during the ascending phase of Solar Cycle 24 of $\sim 7.9\%$, compared to that of $\sim 17\%$ for FISM2. This is largely attributable to FISM2 consistently overestimating TIMED/SEE in the chosen band, which was accounted for in Y_{final} by the fitting of δ.

Final Model Residuals

We observed residuals clustered around zero, but also growing in variance as a function of time progressing through the test set.



(a) Residuals between Y_{final} and TIMED/SEE for final predictions and the associated 95% confidence intervals the bottom of Figure 2.



(b) Distribution of residuals between Y_{final} and TIMED/SEE.

Figure 3. Behavior of the residuals between Y_{final} and TIMED/SEE for the test set (ascending phase of Solar Cycle 24).

The distribution of residuals was leptokurtic, with a kurtosis of $\sim 3.06\%$. The distribution additionally was left-skewed, with a skew of approximately -0.83, indicating that when deviating, Y_{final} was more likely to overestimate than underestimate SEE. The residuals for Y_{final} as well as its associated confidence intervals all demonstrated a downward linear trend. The parameters of each fit to the residuals are in Table 1 below.

Case	Slope	Intercept
Y_{final} (upper)	-3.962×10^{-13}	5.056×10^{-4}
Y_{final}	-1.051×10^{-13}	1.277×10^{-4}
Y_{final} (lower)	-3.957×10^{-13}	4.227×10^{-4}

Table 1. Fitted parameters (rounded to three decimal places) for linear fits to final residuals for Y_{final}.

Conclusions and Future Work

GAMs are a versatile tool allowing for flexibility in the functional form of models fit to time varying data. This research demonstrates that combined with multiple regression to determine a correction term using residuals, they can serve to capture the behavior of solar irradiance well, even if they are only fit to a subset of historical data. The approach, however, is not without concerns. The leptokurtic nature of residuals results from model performance being coupled to underlying fitted data - biases are more likely to exert influence when fitting is done with smaller or arbitrarily-chosen subsets. Bias as a function of wavelength band should also be investigated, as well as the efficacy of the approach when leveraging outputs from other irradiance models such as EUVAC/HEUVAC. Additionally, the issue of forecasting model inputs in a statistically robust manner is vital for operational capability. To this end, future work will involve the evaluation of a coupled approach involving autoregressive models and Fourier analysis.

References

[1] Phillip C Chamberlin, Francis G Eparvier, Victoria Knoer, H Leise, Alicia Pankratz, Martin Snow, Brian Templeman, Edward Michael Benjamin Thiemann, Donald L Woodraska, and Thomas N Woods. The flare irradiance spectral model-version 2 (fism2). *Space Weather*, 18(12):e2020SW002588, 2020.

[2] T Hastie and R Tibshirani. Generalized additive models. chapman hall & crc. *Monographs on Statistics & Applied Probability. Chapman and Hall/CRC*, 1, 1990.