

# Learning and estimating whole sky visible, VNIR, SWIR radiance distributions from a commercial camera

Joseph Del Rocco<sup>a</sup>, Charles Brandon Patterson<sup>b</sup>, Hassen Dhrif<sup>c</sup>, and Joseph T. Kider Jr.<sup>a</sup>

<sup>a</sup>Institute for Simulation and Training, University of Central Florida, Orlando, FL, USA

<sup>b</sup>Full Sail University, Winter Park, FL, USA

<sup>c</sup>University of Miami, Coral Gables, FL, USA

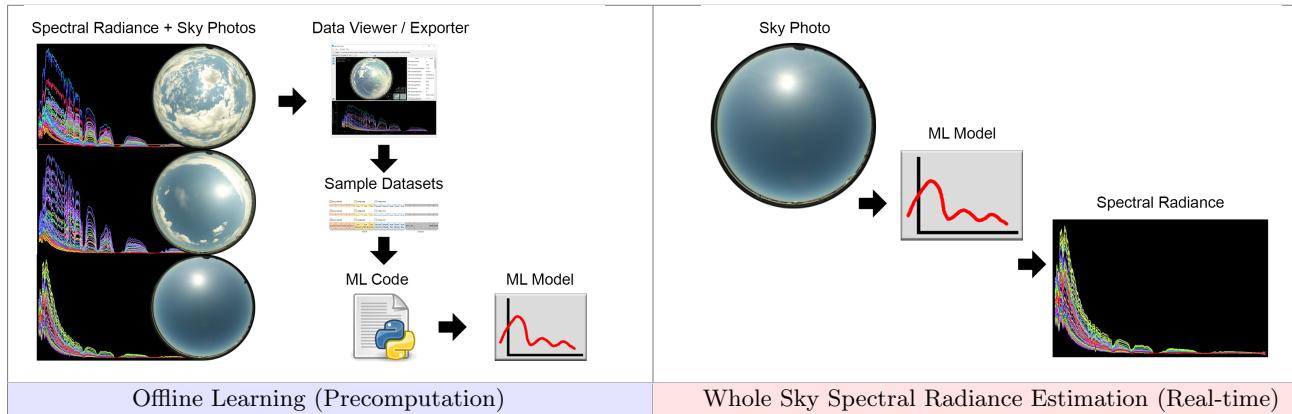


Figure 1. We estimate sky radiance distribution curves between 350-2500nm from images captured with a digital camera. (Left) We use measurements from a commercial digital camera and a sky scanning spectroradiometer to train machine learning algorithms. (Right) We then utilize this ML model to take new sky images and produce the whole sky spectral illumination in real-time. We have validated this approach to measured data.

## ABSTRACT

This work proposes estimating sky radiance distribution curves between 350-2500nm from images captured with a hemispherical digital camera. A novel hardware system simultaneously captured spectral, spatial, and temporal information to acquire accurate physical measurements of the solar/skydome radiance variation. To achieve this goal, we use a custom-built spectral radiance measurement scanner to measure the directional spectral radiance, a pyranometer to measure the irradiance of the entire hemisphere, and a commercial digital camera to capture high-dynamic range (HDR) hemispherical imagery of the sky. We use the measurements obtained from a commercially available digital camera and correlating spectroradiometer measurements to train machine learning (ML) models to estimate whole sky full-spectrum radiance distributions (VIS, VNIR, and SWIR) from a low dimensional RGB input. We train clear, cloudy, and mixed sky models, and cross-validate the estimated radiance distributions with ground-truth data. We highlight important measured and engineered ML features, and we present useful feature engineering techniques employed to minimize model estimation error. Additional contributions of this work include the code for all ML models and experiments, a dataset of all-sky HDR captures with correlating spectroradiometer measurements captured 453 times over 16 days, and an open-source, cross-platform, interactive viewer used to visualize photometric and radiometric data side by side.

**Keywords:** Sky Radiance, Radiance Distribution, Multispectral, Machine Learning, Regression, Sky Viewer

---

Send correspondence to: Joseph Del Rocco, jdelrocco@ist.ucf.edu

## 1. INTRODUCTION

Multispectral sky radiance is needed for wavelength dependent computations in a wide variety of domains and applications, including: radiative transfer,<sup>1</sup> building-performance simulation,<sup>2,3</sup> photovoltaic (PV) panel alignment,<sup>4</sup> climate science,<sup>5</sup> physically based rendering<sup>6,7</sup> natural light transport,<sup>8,9</sup> and any thermal/lighting application using bidirectional reflectance distribution functions (BRDFs).<sup>10-13</sup> The spectra ranges of interest can vary based on a number of factors; we use the following: visible (VIS) (380-780nm), visible and near-infrared (VNIR) (380-1400nm), and short wave infrared (SWIR) (1000-2500nm). Of even more importance is the need for collections of whole sky angular dependent radiance distribution curves required for accurate computations on surfaces affected by the angle of incidence, e.g. walls, facades, PV panels, all of which require more than just a single downward-welling radiance measurement, and all of which need to account for occlusions of the visible skydome.<sup>14</sup> Despite the fact that skylight itself has been studied for well over one hundred years (going back to Lord Rayleigh,<sup>15</sup> Mie,<sup>16</sup> Kimball,<sup>17</sup> and Pokrowski,<sup>18</sup> to name a few), the process of actively measuring or even computing whole sky radiance distributions quickly and accurately still remains an open problem, especially if the curves are to be used in a real-time, practical setting. Reliable, fast, cost-effective methods are still desired. This work attempts to use various modern machine learning methods to train models with the express purpose of being used in a real-time setting to estimate whole sky radiance curves quickly, within some acceptable error, using a dataset of whole sky images and radiance distributions measured with a custom-built hardware system.<sup>19</sup>

Saito et al. estimated sky radiance distributions with a derived equation taking total ozone column and RAW RGB counts.<sup>20</sup> They specifically attempt sky radiance estimation “*without any training sets*,”<sup>20</sup> but focus on a single point in the sky (the zenith) for a subset of visible wavelengths (430-680nm). An important contribution is their camera-specific color matching functions (CMFs), an extension of work done by Sigernes et al.,<sup>21</sup> which take into account camera lens wavelength dependence, vignetting effect, and CMOS noise.<sup>20</sup>

Tohsing et al. used a machine learning approach to estimate full visible spectrum whole sky radiance distributions. They used non-linear regressions, one per color component (RGB) per wavelength (380-760nm), or 1143 models, using the non-linear relationships between RGB counts and radiance values.<sup>22</sup> Although they use 113 sampled points of the hemisphere, their measurements are limited to 8 of 12 consecutive days of the year, with only 1 day’s clear sky data and 1 day’s overcast data used for training.<sup>22</sup> They classify sky cover procedurally into 2 groups (clear or cloudy) via sky index,<sup>23,24</sup> and train 2 separate groups of regressions; scattered skies thus use one or the other. Because full hemisphere measurements took 12 minutes to complete, a synchronized, synthetic image was constructed and sampled.

Chauvin et al. also used a custom-built sky viewer/imager to obtain and then remove/clean anisotropy from clear sky portions of sky images for better cloud detection and ultimately better solar plant control.<sup>25</sup> As stated by the authors, they focus on irradiance/intensity as opposed to radiance curves. Conversely to many sky models, they build upon the work of Grether et al. and Buie et al., noting the importance of the circumsolar region and its ratios along with the central angle from sun position to point position, sun-to-point angle (SPA), and explain how it can be used to estimate clear sky radiance.<sup>25</sup>

Cazorla, Olmo, López-Álvarez, and Alados-Arboledas, et al. used a variety of machine learning techniques, including artificial neural networks (ANN), genetic algorithms (GA) and pseudoinverse linear regression, in a series of projects using images from their own custom-built sky viewer.<sup>5,26,27</sup> The ANN was used to perform cloud detection using color features extracted from sky images while the GA was used to optimize/minimize these features for the input layer of the network.<sup>27</sup> A pseudoinverse linear regression model was used on a datatset of 902 samples of sky image RGB values and visible spectrum radiance at 45°, 60°, and 90° zenith angles. The final training set is whittled down to 40 samples.<sup>27</sup>

Kocifaj’s model, while well researched and presented, might be too computationally expensive for a real-time application, and is therefore tangential to this project.<sup>28,29</sup> Cordero et al. studied albedo effect on radiance distributions (both upwelling and downwelling).<sup>30</sup> Lee studied overcast skies to find meridional consistencies.<sup>31</sup> Pissulla et al. compare 5 separate spectroradiometers and present their measurement deviations.<sup>32</sup>

A discussion on sky radiance distributions is not possible without acknowledging turbidity and its attenuating effects on terrestrial solar radiance. Rayleigh,<sup>15</sup> Mie,<sup>16,33,34</sup> and non-selective scattering algorithms explain it. Although it must be accounted for in models of highest accuracy, it can sometimes be ignored or roughly

estimated. There is a wealth of literature in the atmospheric and solar energy communities that highlight the consistency of the “cosine factor” of solar zenith angle (SZA) on radiance curves, starting with Steven and Unsworth in the mid-1970s, who recorded standard radiance distributions of clear skies, and in doing so found “departures due to variation in atmospheric turbidity [...] to be small”.<sup>35</sup>

## 2. DATA CAPTURE

### 2.1 Hardware

We utilized the capture setup proposed by Kider et al. to capture data needed to train and predict sky radiance distribution curves for the sun/skydome.<sup>19</sup> This novel hardware system captured spectral, spatial, and temporal information, including radiance, HDR fisheye imagery, and total irradiance, simultaneously for various sky conditions. Fig. 2(left) shows the capture setup on the roof of a building. The system includes a custom-built spectral radiance measurement scanner to measure the directional spectral radiance, using an ASD FieldSpec Pro spectroradiometer to capture field measures of the solar spectrum between 350-2500nm (Fig. 2(a)). The system used a commercial digital camera to capture HDR hemispherical imagery of the sky, using a Canon 5D with a Sigma 8mm circular fish-eye lens, following a similar scanning technique to Stumpfel et al.<sup>36</sup> This approach utilized 8 exposures to span a range of 17 stops of the sun and sky to capture a wide dynamic range (Fig. 2(b)). Lastly, the system used an Apogee pyranometer to measure a total irradiance value of the sun/sky (Fig. 2(c)).

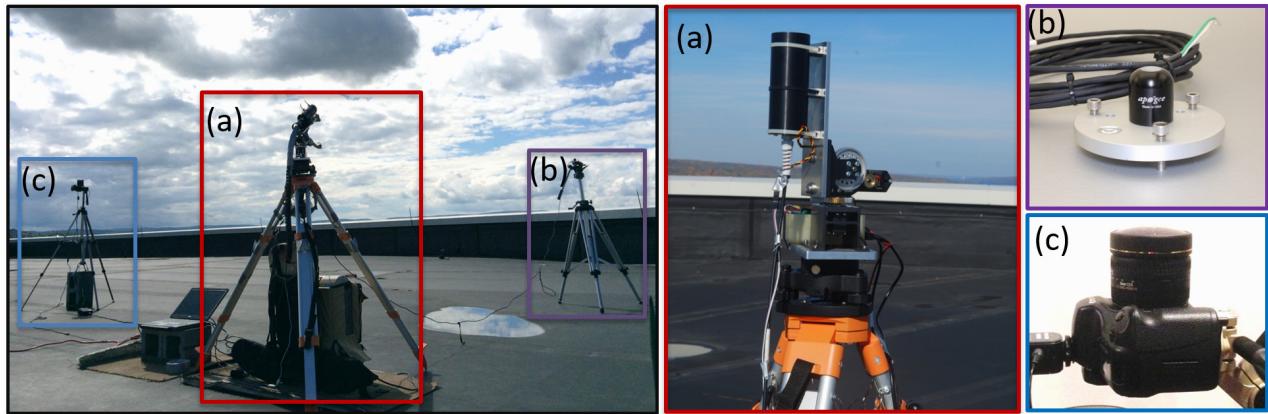


Figure 2. Hardware setup of different capture devices from Kider et al.<sup>19</sup> used to capture data for this project. (a) A custom-built radiance scanner measures spectral radiance between 350-2500nm using a 1° fore-optic lens; (b) Apogee pyranometer, which measures the irradiance of the entire sun/sky; (c) Canon 5D commercial camera with Sigma 8mm circular fish-eye lens captures HDR imagery of the sun and skydome.

### 2.2 Measurements

Throughout this work, hemispherical sky coordinates are specified as *(azimuth, altitude)*, where *altitude* is the complement of *zenith* (90°- *zenith*), and *azimuth* is a horizontal angle increasing eastward from north. Note that our sky photos have north and south inverted due to hardware orientation during capture.

Sky photos were captured at a resolution of 4368x2912 pixels in both JPG and CR2 (Canon Raw) formats in an sRGB color space (one of the highest pixel resolutions for custom-built sky viewers to date). As shown in Fig. 3, 8 separate photos were taken in sequence with exposures 1/8000s, 1/1000s, 1/250s, 1/15s, 1/32s, 1/4s, 1s, 2s, and f-stops f/16 and f/4. Radiance was measured in  $W/m^2/sr/nm$  for 81 evenly distributed 1° solid angles across the sky. Our spectroradiometer produced a single uniform curve from 350-2500nm, containing VIS (380-760nm), VNIR (380-1400nm), and SWIR (1000-2500nm) spectra.

We used the National Renewable Energy Laboratory’s (NREL) Solar Position Algorithm (SPA)<sup>37</sup> to accurately compute the sun’s hemispherical coordinates and path. Some have claimed NREL’s SPA algorithm to be too computationally expensive for real-time applications, recommending the SG2 algorithm instead,<sup>38,25</sup> but we found it to be fast and reliable.

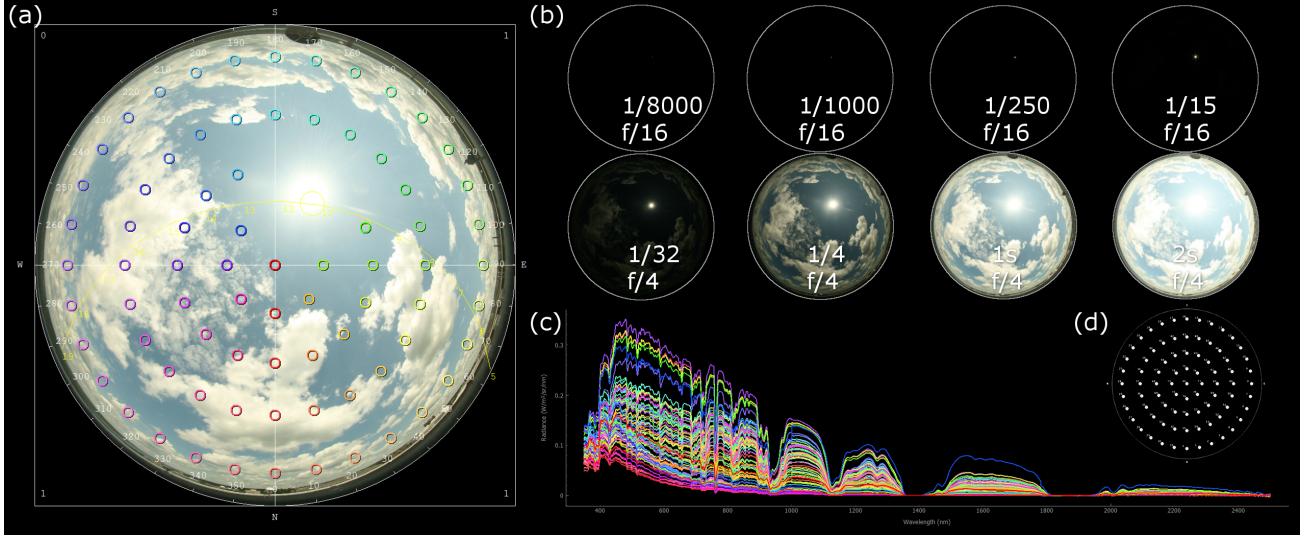


Figure 3. Each data capture includes: (a) a hemispherical photo of the sky, at (b) 8 separate exposures (HDR), and (c) corresponding sky radiance curves, measured at 81 points in the sky with sampling pattern (d). Applying  $20^\circ$  circumsolar avoidance, only 76 of the 81 points of measurement are shown in (a).

### 2.3 Data

All data for this work was captured between 2012-2013 from the rooftop of Frank Rhodes Hall, Cornell University, Ithaca, New York, United States, decimal degrees: (42.443441, -76.481638). Although we measured data well over 1000 times over 41 days throughout the year, with multiple projects in mind, our vetted subset for this work consists of 453 captures over 16 days at differing times of day, covering many different solar angles, all four seasons, and in general a wide variety of atmospheric conditions. These measurements are listed in Table 1, and hosted online for the public’s benefit.\*

As noted in Sec. 2.2, each data capture consists of 8 sky photos (of varying exposure) and 81 radiance curves. A single training/testing sample consists of information extracted from one of these sky photos along with its corresponding directional radiance measurement. Of our 453 total captures intended to be used for this work, 22996 of the 36693 ( $453 \times 81$ ) samples were finally used for training and cross-validation testing. At least 3000+ samples were culled by circumsolar avoidance to avoid the possibility of oversaturation of direct sunlight on our samples, a common approach in many sky models,<sup>25</sup> and important for our RGB-to-curve mapping. We used the same  $20^\circ$  region as Tohsing et al.<sup>22</sup> Note the effects of stray light are drastically reduced at an angle of  $20^\circ$ .<sup>20</sup> Many of the remaining samples culled from our final datasets are due in some part to careful examination of the measurements. This was done with an open-source, cross platform application we developed to load and visualize the measurements.<sup>†</sup> Although a supervised process, it allowed us to quickly navigate the abundance of data to identify dropped, locked, missing, obscured, or oversaturated measurements.

Regarding sky cover, although automatic/procedural assessment is certainly possible,<sup>22–24, 27, 39, 40</sup> we manually assessed our images to ensure accuracy during machine learning. Efficient algorithms can and should be used to automatically assess sky cover during real-time application of this work, for determining which trained model to use. Like Lee,<sup>31</sup> we use the standard categorization of sky conditions provided by the US National Oceanic and Atmospheric Administration (NOAA).<sup>41</sup> We focus on the following three: clear (CLR), scattered (SCT), and overcast (OVC), where CLR and OVC represent 0 and 8 oktas of cloud cover, respectively.<sup>41</sup> In this work, SCT is used to describe skies with any cloud coverage between 1-7 oktas. We ignore the distinction of few (FEW) and broken (BKN) skies to simplify the number of machine learning models that might be used in a practical setting. In theory, our models can be trained on any sky condition.

\*<https://spectralskylight.github.io/RadianceEstimationData>

†<https://spectralskylight.github.io/SkyDataViewer>

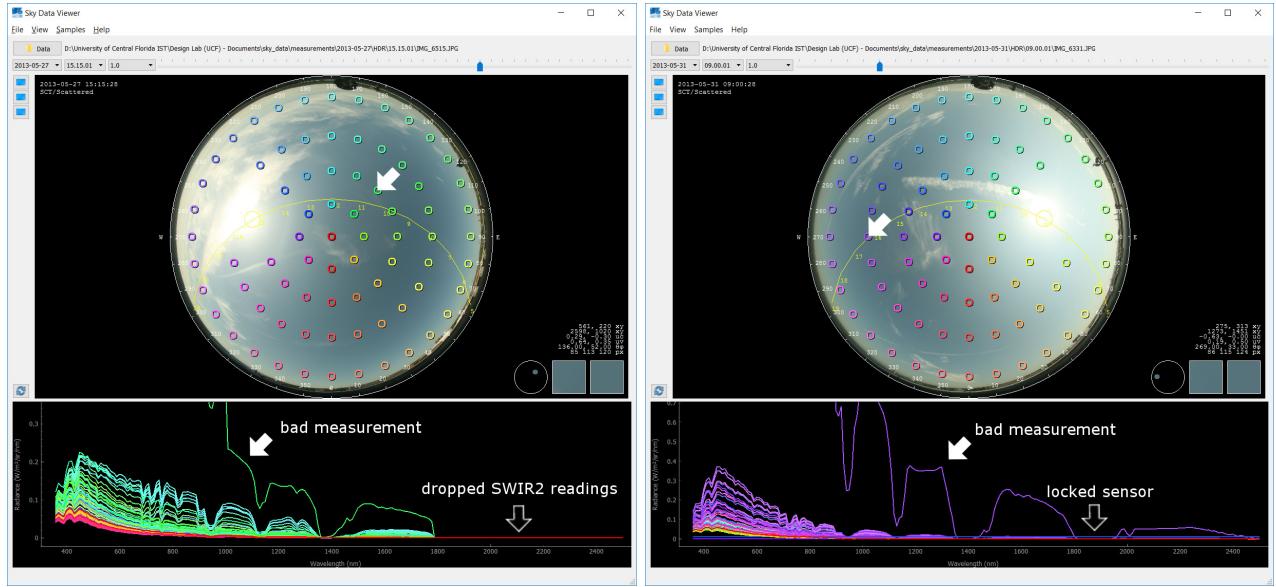


Figure 4. Capturing data *in the wild* for machine learning algorithms is a difficult process. Errors in data capture can vastly affect ML training. Therefore, we developed a SkyDataViewer to help identify data anomalies.

Table 1. List of observed data captures, which include correlated HDR images and radiance distributions.

Date	Time	Captures	Samples	Sky Cover(s)
11/06/2012	12:26 - 16:21	41	3321	SCT
11/15/2012	11:15 - 16:26	56	4536	CLR, SCT
04/13/2013	09:55 - 10:01	2	162	OVC
04/14/2013	10:42 - 18:36	46	3726	SCT, OVC
04/15/2013	07:38 - 08:03	8	648	SCT, OVC
05/12/2013	10:30 - 13:45	14	1134	SCT, OVC
05/26/2013	10:30 - 17:30	29	2349	CLR, SCT
05/27/2013	09:30 - 18:30	37	2997	CLR, SCT
05/30/2013	09:30 - 12:45	14	1134	SCT
05/31/2013	09:00 - 15:00	25	2025	SCT
06/15/2013	07:45 - 18:30	44	3564	CLR, SCT
07/26/2013	11:00 - 14:45	16	1296	CLR, SCT
07/29/2013	09:00 - 14:00	21	1701	SCT, OVC
08/30/2013	09:15 - 14:00	18	1458	SCT, OVC
09/24/2013	06:49 - 18:09	38	3078	CLR
09/26/2013	08:30 - 15:40	44	3564	SCT
16 Days		453	36693	

### 3. SPECTRAL SHAPE ESTIMATION

We approached the problem of estimating radiance distribution curves with machine learning (ML), a superset discipline including much of statistics and neural networks. There are a variety of approaches, algorithms, and software frameworks to choose from. For this work, we focused on a small subset, statistical regression models, and used scikit-learn,<sup>42</sup> a powerful, popular open source library written in Python and Cython,<sup>43</sup> providing most standard ML functionality (supervised/unsupervised algorithms,  $k$ -fold cross validation, principal component analysis, imputation, etc.).

We use  $k$ -fold cross-validation (CV), or rotation estimation,<sup>44,45</sup> to get the most out of our dataset and to minimize the variance of prediction accuracy. We used 10-fold CV specifically to have larger subsamples and less risk of over-fitting. 0 samples from final test captures were used for training, to avoid bias and data leakage.<sup>46</sup>

In general there are three primary problem categories in machine learning: classification, regression, and clustering, and three primary learning modes: supervised, unsupervised, and reinforcement. Predicting a radiance distribution, essentially a continuous curve of points, using correlated ground truths, is a supervised regression problem. Thus our first task was clear - we needed to clean, correlate, and export all usable measurements into datasets of input-output mappings (or samples), each of which correspond to a single directional radiance measurement, and each of which would be used for either CV training/testing or final holdout testing. The initial measured (raw) features extracted from our data include: timestamp, pixel colors (*RGB*), measurement location (*azimuth, altitude*), and directional radiance values per wavelength. The use of multiple exposure (HDR) pixel values and sky irradiance were left for future/ongoing work, though it stands to reason that multiple pixel values across multiple images can just as easily be used as features, as a single pixel value.

Next, we determined which parameters were needed for our model. Feature engineering, the process of determining which inputs best represent the correlated outputs, is an important and often time-consuming part of any machine learning project.<sup>47–50</sup> It is not recommended to rely on raw features alone. Too many features may distract an algorithm, while too few may be insufficient to model complex relationships. Investigative exploratory data analysis (EDA) on the raw features gave us insight and preliminary performance metrics, such as the importance of one feature over another, the insignificance of sample azimuth, etc. Fig. 5 shows a subset of EDA techniques employed.

After EDA, we began searching for an appropriate regression model. Though predicting spectral radiance is clearly a regression problem, it was not so clear which regression model should be used. Many common statistical models (for both regression and classification problems) produce a single output, and thus have high-dimensional

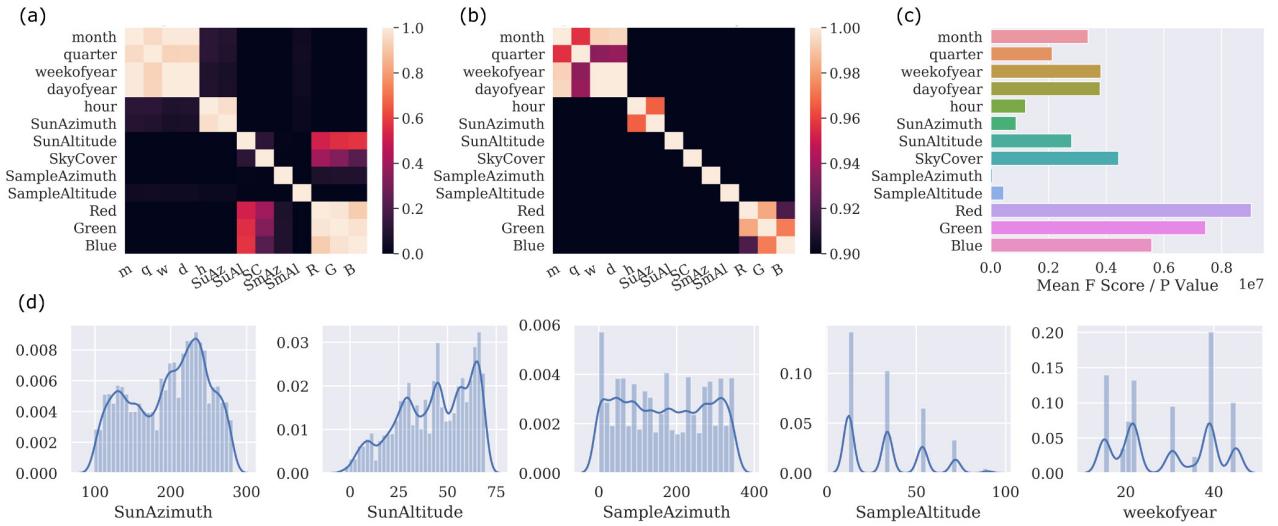


Figure 5. EDA (a) correlation matrix, (b) collinearity matrix, (c) feature importance , and (d) histograms.

input and low-dimensional (i.e. 1) output. We had the exact opposite problem - a handful of input features to predict 2151 values (wavelengths 350-2500nm). Multi-output regression models seemed best suited to the challenge. We started with the simplest possible model as a baseline, the linear regression model (with multi-output support). We refer to this model as LNR. As expected, initial scores were low (since outputs are highly noisy curves, not linear trends), implying under-fitting and the need for more complexity to the inputs.

We accomplished this several ways: (1) by “binning” the capture timestamp into discretized buckets (quarter, month, week, hour, etc.), (2) by computing the sun’s azimuth and zenith/altitude per capture, (3) by tagging samples with sky cover assessment, and (4,5) with polynomial modeling and feature scaling, respectively. In addition to providing a more complex input, method (1) has the benefit of potentially capturing seasonal and daily shifts such as dawn/day/dusk/night, as well as rainy versus arid seasons, all of which have different skies and sometimes trending turbidity.<sup>51</sup> Method (2) complies with the motivation for this work, that the sun’s zenith/altitude might have a large impact on the shape of a radiance curve. The feature importance graph in Fig. 5 confirms this. Summarized by Chauvin et al., when considering radiance at any given point in the sky, of capital importance is the point-to-zenith angle (PZA), the sun-to-zenith angle (SZA), and the central sun-to-point angle (SPA) between them.<sup>25</sup> We believe the importance of this information is captured by both the measured sample position and computed sun position included in each sample. As mentioned in Sec. 2.3, we used NREL’s SPA algorithm for computing sun position per capture.<sup>37</sup> Method (3) is also discussed in Sec. 2.3. Method (4), polynomial modeling, has been used for decades for problems such as curve fitting.<sup>52</sup> We used it to generate a polynomial combination of features from our specified set of inputs along with a polynomial degree (e.g. the set  $[a, b]$  becomes  $[1, a, b, a^2, ab, b^2]$  with a degree of 2).<sup>42</sup> Specifically, we found that using this polynomial modeling technique with a degree of 4 on the inputs of the LNR model, boosted its CV test score considerably. Our final feature engineering contribution, method (5), involved the use of scalers on the inputs: standard, robust, and quantile transformer,<sup>42</sup> all of which attempt to spread the input over a uniform distribution and reduce the impact of marginal outliers.<sup>42</sup> The quantile transformer in particular had the greatest positive impact, and so we applied, saved, and loaded it alongside our models. After data processing and feature engineering, a final train/test sample consists of the input and output features shown in Fig. 6. Future work will include more input features.

We used the following error metrics to compare models and assess prediction performance: coefficient of determination score ( $R^2$ ) and root mean squared deviation (RMSD). We also considered mean bias deviation (MBD), but found it to be “noisy” without much correlation to  $R^2$  and RMSD. We believe this error metric is best suited to a single wavelength, not deviation across a whole curve. The  $R^2$  score we used comes from scikit-learn,<sup>42</sup> and is defined as:

$$R^2(x, y) = 1 - \frac{\sum_{i=1}^N (x_i - y_i)^2}{\sum_{i=1}^N (x_i - \bar{x}_i)^2} , \quad (1)$$

where  $(x, y)$  is a (truth, prediction) pair,  $N$  the number of curves considered, and  $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$ . Note that this score can be negative, despite the name  $R^2$ . As with Tohsing et al.<sup>22</sup> and others, we used the RMSD metric provided by Iqbal,<sup>53</sup> which indicates the variation of predicted from measured, and is defined as:

$$RMSD = \sqrt{\frac{\sum_{i=1}^N (y_i - x_i)^2}{N}} \quad (2)$$

where  $N$  is the number of curves considered,  $x$  the measured/truth curves, and  $y$  the predicted.

						Inputs		Outputs									
						Measured			Computed			Engineered					
						Sample Altitude	Sky Cover	Sun Altitude	Sun Azimuth	Day	Week	Month	Quarter				
quarter	month	week	day	hour													
					Sun Azimuth	Sun Altitude	Sky Cover	Sample Azimuth	Sample Altitude	Pixel Red	Pixel Green	Pixel Blue	350	351	...	2499	2500

Figure 6. A dataset sample consists of 13 input features and 2151 output features (the radiance curve).

### 3.1 Machine Learning Models

After feature engineering and preliminary tests with our polynomial-input linear regression model (LNR), we investigated other multi-output regression models, such as Lasso,<sup>54</sup> Ridge,<sup>55</sup> ElasticNet,<sup>56</sup> and Lars. The Lasso linear estimator model has a built-in regularizer, which removes outlier features with low correlation to the output, in an attempt to reduce over-fitting. Ridge is similar, except instead of removing the outliers, it artificially penalizes them to reduce their effect. ElasticNet seems to be a combination of the two. Lars is similar to stepwise regression, in that it finds the most correlated predictors. Though it is typically effective with high dimensional data, it is also sensitive to noise. Unfortunately, as it turned out, all of these models performed no better than LNR, our initial polynomial-input linear regression model, perhaps due to the highly noisy nature of radiance curves, or the fact that we were not over-fitting the data.

The  $k$ -NN regressor ( $k$ -nearest-neighbors or  $k$ -neighbors) (KNR) model first gathers all data and then derives predictions from local interpolation of the  $k$ -closest data points. KNR uses a spatial partitioning tree to organize the data for fast searching and filtering of data points. KNR was the first model we found that performed better than LNR.

The final model(s) experimented with that performed well were decision tree regressors. Decision tree regression models approximate curves by employing a set of “if-then-else” rules to predict. The tree depth hyperparameter determines how many rules are used in the tree. If a decision tree regressor becomes too deep, the regressor essentially overfits the data, as the deep chain of rules cannot generalize well. The random forest regressor (RFR)<sup>57</sup> and extra tree regressor (ETR)<sup>58</sup> both address over-fitting with many separate internal trees for different portions of the dataset, and since each tree is limited in scope, it is not able to overfit. Once final models were chosen, additional hyperparameter tuning was done. Parameters of an ML model are often determined procedurally during training, such as weights and coefficients, whereas hyperparameters are typically specified by the user, such as tree depth or number of trees, etc.

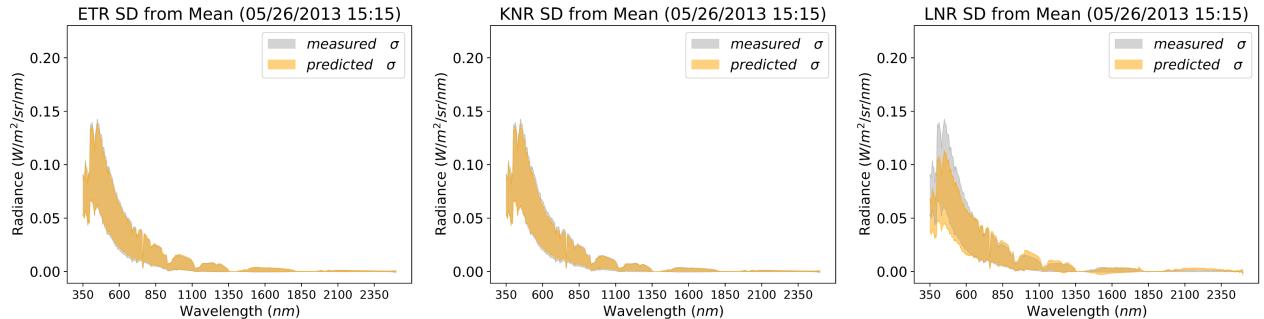


Figure 7. Three of the four final regression models, ETR, KNR, and LNR, respectively, showing measured and predicted standard deviation from mean on test capture 05/26/2013 15:15 (clear sky). RFR performed very similar to ETR.

Table 2. Holdout error for final regression models using 10-fold cross validation training and testing over *all usable test samples, across all skies* with train-test ratio of 3:1.<sup>§</sup>

ID	Name	R <sup>2</sup>	RMSD
ETR	Extra Trees Regressor	66.55%	82.62%
RFR	Random Forest Regressor	62.84%	87.09%
KNR	K Neighbors Regressor	58.57%	91.95%
LNR	Linear Regressor	51.02%	99.99%

<sup>§</sup>The high error was mostly due to scattered sky cover predictions, as shown in Sec. 4

## 3.2 Sky Cover Models

After cross-validated training and testing of our conglomerate (“mix”) dataset, with all usable training samples across all sky covers, we suspected that models trained on sky specific datasets might prove more effective. The feature importance graph in Fig. 5 confirms this. Thus we filtered the mix dataset into 3 additional datasets (clear, scattered, overcast) based on sky cover assessment (CLR, SCT, OVC, respectively), and divided each of them for CV training/testing and holdout validation testing with our final regression models. The mix dataset contained all usable samples and was almost 1GB in size. Training and testing on a smaller, more focused dataset should be computationally cheaper and possibly even more precise, since the training data better reflects the expected real-world data. As discussed in Sec. 2.3, there are a variety of automated/procedural sky cover algorithms that could be employed in a real-time system to assess the input sky images to quickly determine which trained model to use for prediction. One problem with this experiment was that clear, scattered, and overcast datasets contained 5548, 15508, and 1940 samples respectively (i.e. 24%, 67%, and 8% of total usable samples), and thus overcast skies were not well represented.

## 4. RESULTS AND DISCUSSION

### 4.1 Regression Models

As the results in Fig. 8 show, when models are trained on all training data across all skies, ETR generally outperforms the rest, with RFR and KNR trading places depending upon the capture. More tests are necessary, but one strategy for real-time application of this work might be triple-mode redundancy,<sup>59</sup> or “democratic computing,” using multiple regressors to predict with (since they are computationally efficient), and using the predicted curves from a specific model based on some decision factor (e.g. standard deviation from mean, etc.)

Although the LNR model was by far the worst performing, we suspect it performed better than expected on some captures (clear skies, possibly with low turbidity) because of the generally similar shape to radiance distributions. As seen in plots showing all measured radiance distributions across a clear sky, as well as time-series plots, the shape of radiance curves can be relatively consistent, often times varying only in magnitude. In this way, each of the coefficients of the polynomial-input LNR model might actually be near-linear as the magnitude of the curve changes. In fact, as Fig. 9 shows, LNR might be usable as-is on some captures (again, as one of several models in a collection of predictors).

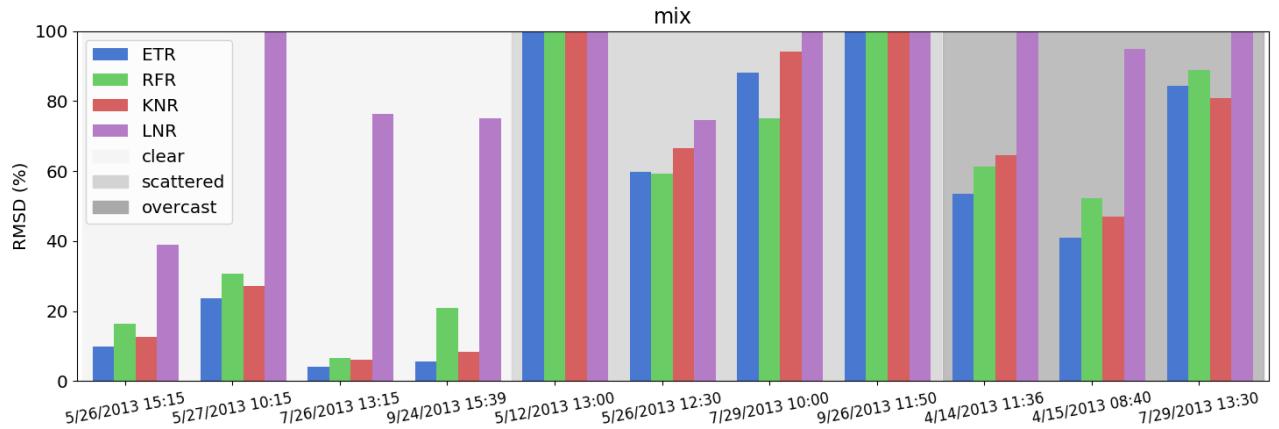


Figure 8. RMSD of the models trained with mix dataset, which includes all usable testing samples across all skies.

### 4.2 Sky Cover

Final results show that sky cover was extremely relevant, as initially suspected. Three of the four final models were very successful in predicting clear sky captures, shown in Tbls. 3 and 4, and Figs. 10, 11, and 12. LNR even predicted well on two clear sky captures, as mentioned in Sec. 4.1.

For scattered sky models, RFR's predictions are a surprising success (discussed below). All other models had a hard time with the data. In general, we posit that scattered skies are not predicted well by our methods (as of yet), despite the very large training set of scattered samples. There are a number of reasons why this could be. First, although lens distortions from curvature, vignetting, and even wavelength-specific variation were known to us from Saito et al.,<sup>20</sup> we hadn't yet compensated for it in this work. Even a slight shift in sampling coordinates could severely alter the RGB values extracted from the sky images. With clear sky data, that may not be very noticeable, as the color of the sky is fairly uniform, but scattered sky clouds can vary considerably. Future work should definitely compensate for it. Second, the pixel region and weighting algorithms we used may not have been optimal for cloudy skies. Again, clear skies exhibit relatively uniform color whereas clouds and cloud/sky borders can vary, which would distort a final pixel value if weighted over a large enough kernel.

Only three overcast skies were held out for validation, and as mentioned in Sec. 3.2, only 8% of the mix dataset were overcast skies (1940 samples). So the results of this experiment on overcast skies are of low confidence. Despite that, ETR, KNR and RFR models were still able to predict one of those three captures with 40% RMSD. Given more data, our methods may prove successful on overcast skies as well.

Of particular interest is RFR's performance trained with a mix dataset versus a sky cover specific dataset. RFR performed worse than ETR and KNR on most validation captures when trained with mix. However, it performed equivalent or better than KNR and ETR when trained with sky cover specific data. In fact, as Fig. 9 shows, RFR even predicted scattered skies with semi-acceptable error, something none of the other models were able to accomplish.

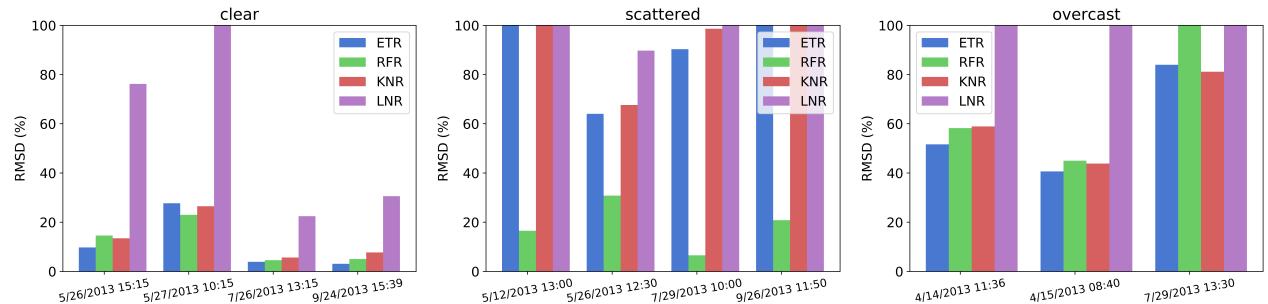


Figure 9. RMSD of the models trained with sky cover specific datasets.

Table 3. Error metrics for final validation clear sky captures predicted with ETR model clear dataset.

Model	Capture	Sky	R <sup>2</sup>	RMSD
ETR	05/26/2013 15:15	CLR	98.31%	9.73%
ETR	05/27/2013 10:15	CLR	84.31%	27.7%
ETR	07/26/2013 13:15	CLR	99.16%	3.88%
ETR	09/24/2013 15:39	CLR	99.55%	3.08%

Table 4. Error metrics of actual vs predicted radiance curves for three select samples shown in Fig. 11 .

Sample Index	Coordinates	R <sup>2</sup>	RMSD
11	(123.75°, 12.1151°)	99.97%	1.95%
27	(315.0°, 12.1151°)	100.00%	0.57%
80	(303.75°, 90°)	99.89%	1.92%

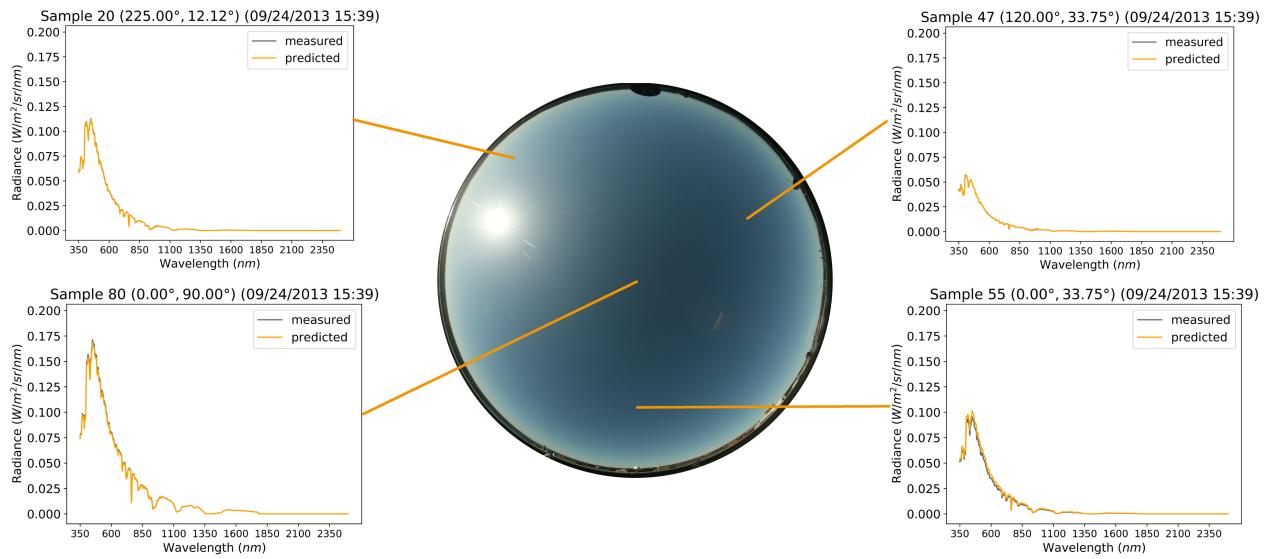


Figure 10. Select samples from clear sky capture 09/24/2013 15:39 predicted with ETR model clear dataset.  
Video 1. <http://dx.doi.org/doi.number.goes.here>

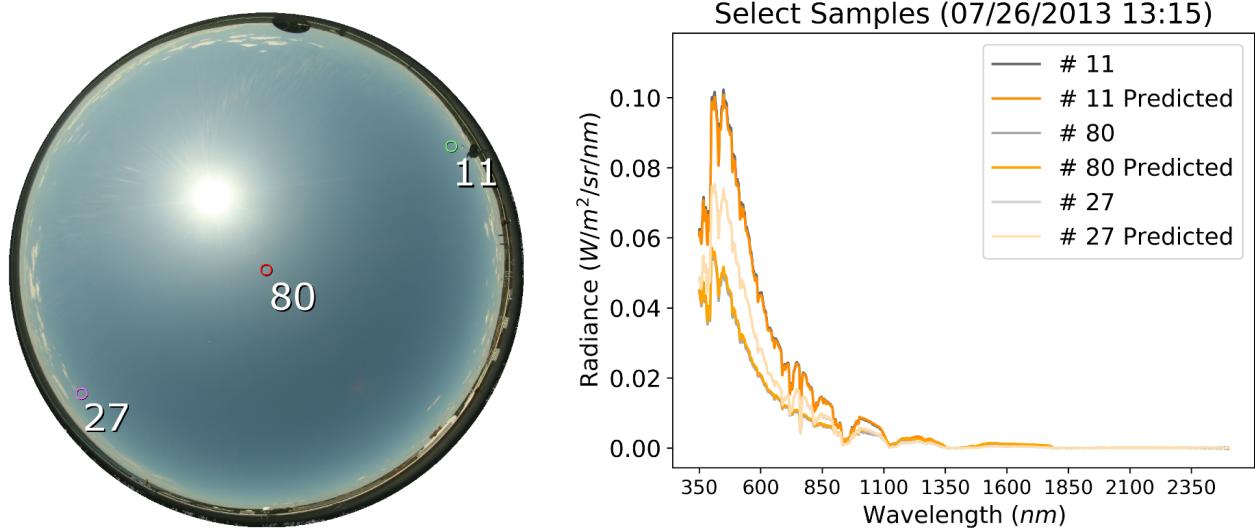


Figure 11. Actual vs predicted radiance distributions for 3 select samples captured on 07/26/2013 13:15. Sample (80) is at the zenith, while (11) and (27) are near the horizon.

### 4.3 Whole Sky

This work attempted to learn and predict collections of whole sky angular dependent radiance distribution curves for real-time application, because of the high demand across multiple disciplines for spectral illumination. Figs. 10, 11, and 12 show that this is certainly possible with clear skies. All four holdout validation captures were predicted with success, three under 10% error.

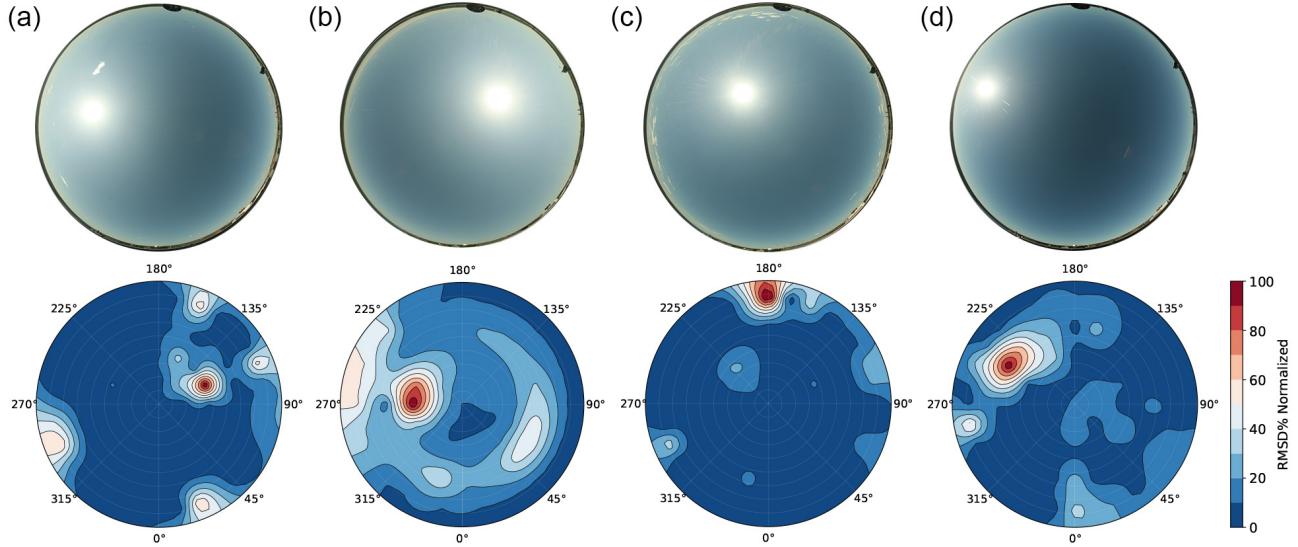


Figure 12. Demonstration of whole sky accuracy from holdout test data using ETR model trained on clear dataset. These captures correlate with overall error reported in Fig. 9 and Tbl. 3. This figure shows normalized RMSD% for captures (a) 05/26/2013 15:15, (b) 05/27/2013 10:15, (c) 07/26/2013 13:15, and (d) 09/24/2013 15:39. We compare the generated spectral data with the captured data. (No samples from these captures were used for training.)

## 5. CONCLUSIONS

We set out to develop a machine learning method for estimating sky radiance distribution curves (including VIS, VNIR, and SWIR spectra) from images captured with a hemispherical commercial digital camera. We feel this project was a success in that regard. There is still much room for future work to greater improve accuracy for cloudy and mixed sky conditions, such as: tuning the RFR model, integrating turbidity and cloud type data from remote sensed sources like the GOES satellites, interpolating radiance curves for any point of a skydome, looking at the optimal non-linear regression for different channels across different color spectra (HLS, HSV, etc.), using PZA, SZA and SPA angles instead of (or in addition to) sky coordinates, taking advantage of our HDR data, investigating many other ML approaches (ANNs, etc.).

We hope this work will set the stage for future interdisciplinary research questions and experiments to help improve building energy simulation, photovoltaic placement, spectral rendering applications, etc. Having an online system that provides real-time whole sky angular spectral radiance distributions will improve real-time control systems.

## ACKNOWLEDGMENTS

We would like to formally acknowledge Dr. Donald P. Greenberg and the Program of Computer Graphics (PCG) at Cornell University for their support during data capture and open source release of the sky data used for analysis in this project; Kevin Pratt, Hurf Sheldon, and Lars Schumann for helping build the sky scanner; Dr. William D. Philpot for many discussions on spectral measurements with the ASD FieldSpec device; Dr. Harold van Es for lending PCG the ASD spectroradiometer to conduct experiments; and Daniel Knowlton for the predecessor to our SkyDataViewer and work on the spectral sky project.

## REFERENCES

- [1] Chandrasekhar, S., [*Radiative transfer*], Courier Corporation (2013).
- [2] Jakica, N., “State-of-the-art review of solar design tools and methods for assessing daylighting and solar potential for building-integrated photovoltaics,” *Renewable and Sustainable Energy Reviews* (2017).
- [3] Hensen, J. L. and Lamberts, R., [*Building performance simulation for design and operation*], Routledge (2012).
- [4] Smith, C. J., Forster, P. M., and Crook, R., “An all-sky radiative transfer method to predict optimal tilt and azimuth angle of a solar collector,” *Solar Energy* **123**, 88–101 (2016).
- [5] López-Álvarez, M. A., Hernández-Andrés, J., Romero, J., Olmo, F. J., Cazorla, A., and Alados-Arboledas, L., “Using a trichromatic CCD camera for spectral skylight estimation,” *Applied optics* **47**(34), H31–H38 (2008).
- [6] Jakob, W., “Mitsuba renderer,” (2010). <http://www.mitsuba-renderer.org>.
- [7] Haber, J., Magnor, M., and Seidel, H.-P., “Physically-based simulation of twilight phenomena,” *ACM Transactions on Graphics (TOG)* **24**(4), 1353–1373 (2005).
- [8] Veach, E. and Guibas, L. J., “Metropolis light transport,” in [*Proceedings of the 24th annual conference on Computer graphics and interactive techniques*], 65–76, ACM Press/Addison-Wesley Publishing Co. (1997).
- [9] Jarosz, W., [*Efficient Monte Carlo methods for light transport in scattering media*], Citeseer (2008).
- [10] Kurt, M. and Edwards, D., “A survey of brdf models for computer graphics,” *ACM SIGGRAPH Computer Graphics* **43**(2), 4 (2009).
- [11] Cook, R. L. and Torrance, K. E., “A reflectance model for computer graphics,” *ACM Transactions on Graphics (TOG)* **1**(1), 7–24 (1982).
- [12] Deering, D. W. and Eck, T. F., “Atmospheric optical depth effects on angular anisotropy of plant canopy reflectance,” *International Journal of Remote Sensing* **8**(6), 893–916 (1987).
- [13] Nicodemus, F. E., “Directional reflectance and emissivity of an opaque surface,” *Applied optics* **4**(7), 767–775 (1965).
- [14] Schumann, L. and Greenberg, D. P., “Environment mapping for fast and robust calculation of indirect radiant energy,” *Proceedings of SimBuild* **5**(1), 263–269 (2012).
- [15] Strutt, J. W., “On the light from the sky, its polarization and colour,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **41**(271), 107–120 (1871).
- [16] Mie, G., “Beiträge zur optik trüber medien, speziell kolloidaler metallösungen,” *Annalen der physik* **330**(3), 377–445 (1908).
- [17] Kimball, H. H., “Intensity of solar radiation at the surface of the earth, and its variations with latitude, altitude, season, and time of day,” *Monthly Weather Review* **63**(1), 1–4 (1935).
- [18] Pokrowski, G., “Über einen scheinbaren mie-effekt und seine mögliche rolle in der atmosphärenoptik,” *Zeitschrift für Physik* **53**(1-2), 67–71 (1929).
- [19] Kider Jr, J. T., Knowlton, D., Newlin, J., Li, Y. K., and Greenberg, D. P., “A framework for the experimental comparison of solar and skydome illumination,” *ACM Transactions on Graphics (TOG)* **33**(6), 180 (2014).
- [20] Saito, M., Iwabuchi, H., and Murata, I., “Estimation of spectral distribution of sky radiance using a commercial digital camera,” *Applied Optics* **55**, 415 (Jan. 2016).
- [21] Sigernes, F., Holmes, J. M., Dyrland, M., Lorentzen, D. A., Svenøe, T., Heia, K., Aso, T., Chernouss, S., and Deehr, C. S., “Sensitivity calibration of digital colour cameras for auroral imaging,” *Optics express* **16**(20), 15623–15632 (2008).
- [22] Tohsing, K., Schrempf, M., Riechelmann, S., and Seckmeyer, G., “Validation of spectral sky radiance derived from all-sky camera images-a case study,” *Atmospheric Measurement Techniques* **7** (2014), Nr. 7 (2014).
- [23] Yamashita, M., Yoshimura, M., and Nakashizuka, T., “Cloud cover estimation using multitemporal hemisphere imageries,” *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences* **35**(7), 826–829 (2004).
- [24] Saito, M. and Iwabuchi, H., “Cloud Discrimination from Sky Images Using a Clear-Sky Index,” *Journal of Atmospheric and Oceanic Technology* **33**, 1583–1595 (Aug. 2016).

- [25] Chauvin, R., Nou, J., Thil, S., and Grieu, S., “Modelling the clear-sky intensity distribution using a sky imager,” *Solar Energy* **119**, 1–17 (Sept. 2015).
- [26] Cazorla, A., Olmo, F., and Alados-Arboledas, L., “Using a sky imager for aerosol characterization,” *Atmospheric Environment* **42**(11), 2739–2745 (2008).
- [27] Cazorla, A., Olmo, F. J., and Alados-Arboledas, L., “Development of a sky imager for cloud cover assessment,” *JOSA A* **25**(1), 29–39 (2008).
- [28] Kocifaj, M., “Angular distribution of scattered radiation under broken cloud arrays: an approximation of successive orders of scattering,” *Solar Energy* **86**(12), 3575–3586 (2012).
- [29] Kocifaj, M., “Unified model of radiance patterns under arbitrary sky conditions,” *Solar Energy* **115**, 40–51 (May 2015).
- [30] Cordero, R. R., Damiani, A., Ferrer, J., Rayas, J., Jorquera, J., Tobar, M., Labbe, F., and Laroze, D., “Downwelling and upwelling radiance distributions sampled under cloudless conditions in Antarctica,” *Applied Optics* **52**, 6287 (Sept. 2013).
- [31] Lee Jr, R. L., “Measuring overcast colors with all-sky imaging,” *Applied optics* **47**(34), H106–H115 (2008).
- [32] Pissulla, D., Seckmeyer, G., Cordero, R. R., Blumthaler, M., Schallhart, B., Webb, A., Kift, R., Smedley, A., Bais, A. F., Kouremeti, N., Cede, A., Herman, J., and Kowalewski, M., “Comparison of atmospheric spectral radiance measurements from five independently calibrated systems,” *Photochemical & Photobiological Sciences* **8**(4), 516 (2009).
- [33] Wriedt, T., “Mie theory: a review,” in [The Mie Theory], 53–71, Springer (2012).
- [34] Hergert, W. and Wriedt, T., [The Mie Theory: basics and applications], vol. 169, Springer (2012).
- [35] Steven, M. D. and Unsworth, M. H., “Standard distributions of clear sky radiance,” *Quarterly Journal of the Royal Meteorological Society* **103**(437), 457–465 (1977).
- [36] Stumpfel, J., Tchou, C., Jones, A., Hawkins, T., Wenger, A., and Debevec, P., “Direct hdr capture of the sun and sky,” in [Proceedings of the 3rd International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa], *AFRIGRAPH '04*, 145–149, ACM, New York, NY, USA (2004).
- [37] Reda, I. and Andreas, A., “Solar position algorithm for solar radiation applications,” *Solar energy* **76**(5), 577–589 (2004).
- [38] Blanc, P. and Wald, L., “The sg2 algorithm for a fast and accurate computation of the position of the sun for multi-decadal time period,” *Solar Energy* **86**(10), 3072–3083 (2012).
- [39] Li, Q., Lu, W., and Yang, J., “A hybrid thresholding algorithm for cloud detection on ground-based color images,” *Journal of atmospheric and oceanic technology* **28**(10), 1286–1296 (2011).
- [40] Arking, A. and Childs, J. D., “Retrieval of cloud cover parameters from multispectral satellite images,” *Journal of Climate and Applied Meteorology* **24**(4), 322–333 (1985).
- [41] Office Of The Federal Coordinator For Meteorological Services And Supporting Research, “Chapter 9: Sky Condition,” in [Federal Meteorological Handbook No. 1: Surface Weather Observations and Reports], 45–50, National Oceanic and Atmospheric Administration, Washington, D.C., fcm-h1-2017 ed. (Nov. 2017).
- [42] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., “Scikit-learn: Machine learning in python,” *Journal of machine learning research* **12**(Oct), 2825–2830 (2011).
- [43] Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D. S., and Smith, K., “Cython: The best of both worlds,” *Computing in Science & Engineering* **13**(2), 31–39 (2011).
- [44] Picard, R. R. and Cook, R. D., “Cross-validation of regression models,” *Journal of the American Statistical Association* **79**(387), 575–583 (1984).
- [45] Kohavi, R. et al., “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in [*Ijcai*], **14**(2), 1137–1145, Montreal, Canada (1995).
- [46] Cawley, G. C. and Talbot, N. L., “On over-fitting in model selection and subsequent selection bias in performance evaluation,” *Journal of Machine Learning Research* **11**(Jul), 2079–2107 (2010).
- [47] Michalski, R. S., Carbonell, J. G., and Mitchell, T. M., [Machine learning: An artificial intelligence approach], Springer Science & Business Media (2013).
- [48] Domingos, P., “A few useful things to know about machine learning,” *Communications of the ACM* **55**(10), 78–87 (2012).

- [49] Forman, G., “An extensive empirical study of feature selection metrics for text classification,” *Journal of machine learning research* **3**(Mar), 1289–1305 (2003).
- [50] Brownlee, J., “Machine learning mastery with python,” *Machine Learning Mastery Pty Ltd*, 100–120 (2016).
- [51] Power, H. C. and Willmott, C. J., “Seasonal and interannual variability in atmospheric turbidity over south africa,” *International Journal of Climatology: A Journal of the Royal Meteorological Society* **21**(5), 579–591 (2001).
- [52] De Boor, C., [*A practical guide to splines*], vol. 27, Springer-Verlag New York (1978).
- [53] Iqbal, M., [*An introduction to solar radiation*], Elsevier (2012).
- [54] Tibshirani, R., “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)* , 267–288 (1996).
- [55] Hoerl, A. E. and Kennard, R. W., “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics* **12**(1), 55–67 (1970).
- [56] Zou, H. and Hastie, T., “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320 (2005).
- [57] Kocev, D., Vens, C., Struyf, J., and Džeroski, S., “Tree ensembles for predicting structured outputs,” *Pattern Recognition* **46**(3), 817–833 (2013).
- [58] Geurts, P., Ernst, D., and Wehenkel, L., “Extremely randomized trees,” *Machine learning* **63**(1), 3–42 (2006).
- [59] Anderson, J., “Multiple redundancy applications in a computer,” in [*Proc. 1967 Annual Symposium on Reliability*], 553–562 (1967).