Predicting Atmospheric Parameters using Canonical Correlation Analysis

Emmett J. Ientilucci
Digital Imaging and Remote Sensing Laboratory
Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology

Table of Contents

Table of Figures

1 Introduction

Of interest to many in the remote sensing community, is the make-up of our atmosphere. When airborne or satellite based imaging sensors capture spectrographic information about a piece of real-estate below, it is usually of prime importance for the remote sensor (scientist) to estimate what the ground leaving parameters were (reflectance or radiance). This task can only be achieved if there is adequate knowledge about the atmosphere, for example. When this information is known, algorithms can be applied to "back-out" the effects of the atmosphere thus leaving spectroscopic corrected imagery as though is was measured from the ground itself.

There are many approaches to solving the above-mentioned problem. One such approach is to use the multivariate statistical technique known as canonical correlation analysis (CCA). It is this approach that will be illustrated in this report. However, before we continue we preclude our CCA approach with more traditional approaches in hopes of establishing a motivational factor for using CCA.

2 Forward Modeling

We first introduce the concept of "forward" modeling. A schematic of this concept can be seen in Figure 1. Here we see many input parameters, such as atmospheric profiles (*e.g.*, transmission, temperature), being feed into a physics based model. The physics based model is a model that is usually based on radiative transfer physics of how light energy propagates and interacts with media in the atmosphere, for example (*e.g.*, CO_2, O_3, etc.). However, the model can be a predictor for just about anything (*e.g.*, in-water parameters), other than the atmospheric parameters presented here.

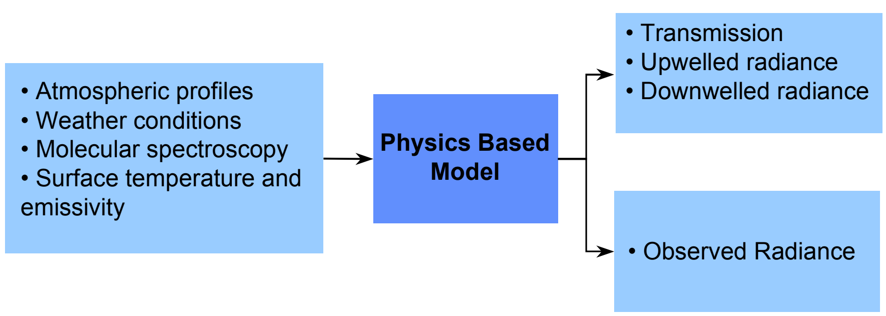


Figure 1 Concept of forward modeling.

The output of the model yields parameters such as atmospheric transmission, upwelled and downwelled radiance, and overall observed radiance. It is the observed radiance that is actually measured by the airborne sensor. It is this parameter that we wish to invert in order to estimate atmospheric profiles, for example.

Before attempting to create an invertible model, we can take the approach of actually measuring the atmospheric profile during the time of image acquisition. This information can be obtained from radiosonde measurements. Radiosondes are balloon-borne sensors launched at approximately the same time and place as the image acquisition. The radiosondes measure temperature, pressure, wind speed, and humidity during the balloon's ascent. This results in vertical profiles that can be used to model the optical properties of the intervening atmosphere between the sensor and the ground. Unfortunately, radiosondes are susceptible to drift during their ascent and may not accurately represent the actual composition of the atmosphere for a given column of air. Furthermore, the logistics of successfully launching a coincident radiosonde for every remote sensing acquisition over the planet is impractical.

2.1 Model Matching

One approach that uses the forward model without ancillary data is to dynamically change the input parameters until the difference between the model output and the observed radiance is minimized based on some criterion (*e.g.*, least squares). This is known as the *model-matching* approach (see Figure 2). Inputs using this approach can be pressure depth, column water vapor, aerosol amount, concentration of chlorophyll, suspended materials or dissolved organics in a water column.

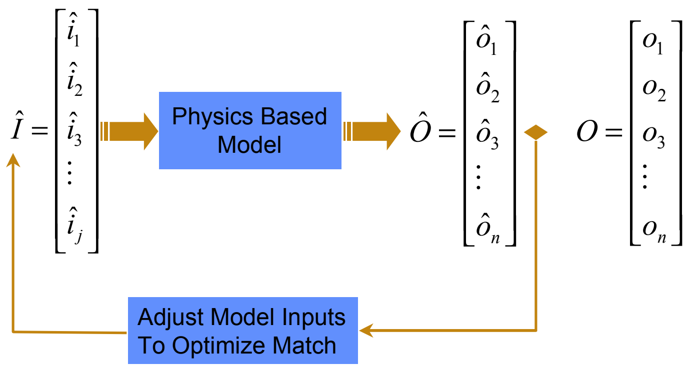


Figure 2 Model matching concept with inputs,($\hat{\imath}$), outputs($\hat{O}$), and measured outputs (O).

The output here is a spectral reflectance or radiance. Model matching requires a good parameterization of the inputs that must be developed *a priori* for each specific application. Also, unless properly constrained, this approach may not converge or may lead to unrealistic atmospheric parameters. The solutions are obtained by iterations that are often nonlinear. Finally, the atmospheric conditions used to match the observed spectra are then used in the radiative transfer equation to solve for the surface radiance.

2.2 Statistical Optimization Methods

Another approach is to use a statistical approach to estimate relationships between input and output parameters. This approach can be seen in Figure 3. This can be thought of as an extension to the model matching approach where we introduce statistical optimization methods driven by model-based data to solve simple linear relationships.

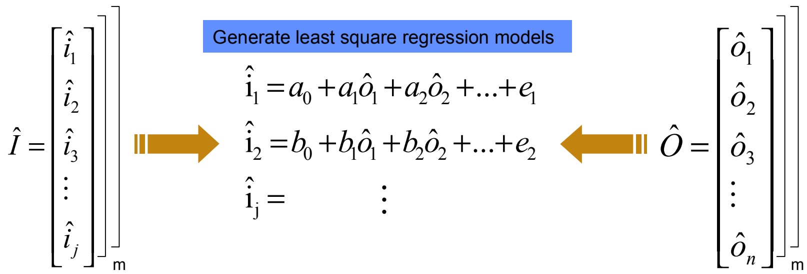


Figure 3 Simple model based regression approach relating input and output parameters. a_0, a_1, b_0, b_1, are the regression coefficients while e_1, e_2 are the errors associated with the regression model.

Here we generate *m* estimates of observed values (Ô) using *m* sets of inputs (î) to the physics based model. This approach typically works well only if the ground measurements (*i.e.*, i_1, i_2, etc.) are individually highly correlated with some subset of the entries in the observed vector.

2.2.1 Canonical Correlation (Regression) Analysis

When we have the situation where the observed values (*e.g.,* spectral radiance) are a function of *multiple* unconstrained model input parameters (*e.g*., target temperature, atmospheric temperature profile, etc.), we need to generate a multi-parameter statistical model that relates predicted observed vectors, to model input parameter vectors (see Figure 4). The multi-parameter model we will use is the use of canonical correlation analysis in a regression scheme. Even more important is the invertibility of the physics based model. More than likely, the physics based model will be difficult, if not impossible, to invert. It would behoove us to generate another model that reasonably approximates the physics based model *and* is invertible.

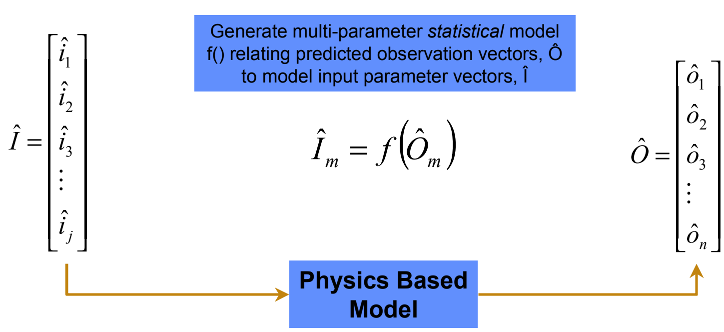


Figure 4 Illustration of multi-parameter model (function) relating input and output parameters.

The CCA method attempts to generate an optimum statistical relationship between a number of input variables (one or more of which are generally the parameters of interest) and a number of output variables (usually the observed spectral vectors) through a joint covariance matrix. The relationship is linear and invertible such that we can invert radiance spectra to desired input parameters. A diagram illustrating the relationships between variables can be seen in Figure 5.

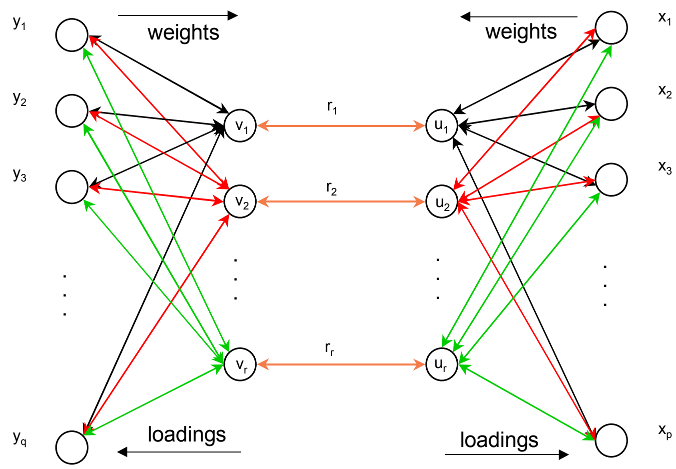


Figure 5 Relationship between input and output variables in CCA.

One very important aspect of the linear combinations in CCA is that they are orthogonal. This ensures that the information about the parameter of interest exploited by one linear combination is linearly independent of the information carried by the other linear combinations. By doing this, CCA avoids redundant information and is able to find the minimal dimension required to adequately predict the parameters of interest. Further investigation and theoretical development of CCA is omitted in this report, for the reports main goal is on the application of CCA.

3 CCA Implementation

To do the analysis, the CCA covariance matrices were built with a suitable training set. That is, we need to train the model where the outcome will be a matrix of regression coefficients relating the canonical variables U and V. The data sets (X and Y) used for the training were created using an atmospheric propagation model called MODTRAN. Given a set of input parameters (*e.g.*, atmospheric pressure, temperature, humidity, ozone, etc.), the physics based model (MODTRAN, in this case) can predict output parameters such as atmospheric transmission, upwelled and downwelled radiance, and overall sensor reaching observed radiance (up to 100km). Figure 6 illustrates the overall approach to training and estimation using the matrix of regression coefficients, β_{cc}.

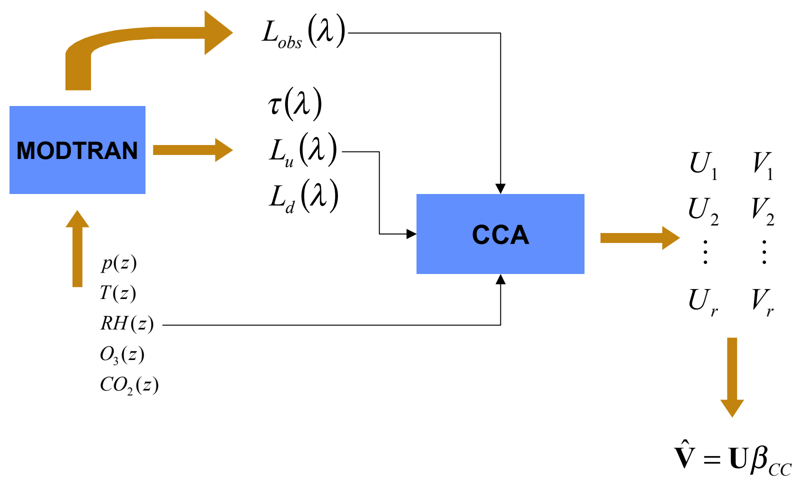


Figure 6 Implementation of CCA to predict atmospheric profiles from observed radiance values.

In order to generate a robust enough model (β_{cc}), a mixture of realistic input atmospheric parameter profiles were synthetically generated and used as inputs to the physics based model (MODTRAN). The end result was 216 runs of MODTRAN which generated 216 observation pairs of input and output parameters. These pairs were then used as training data to CCA which generated a matrix of regression coefficients, β_{cc}.

The input atmospheric parameter profiles were generated with a 3-factor experimental design with no repeats. The factors were the vertical temperature profile, the vertical relative humidity profile, and the total amount of ozone, (which were 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0 times the default value for the input model atmosphere). There were six different temperature and humidity profiles. There were also six different levels of ozone. These vertical atmospheric profiles, along with that of carbon dioxide and pressure, can be seen in Figure 7 - Figure 9. Because the pressure and carbon dioxide do not vary much, these were not a controlled factor. The temperature and humidity profiles used were the default profiles for the six model atmospheres included in MODTRAN. In MODTRAN these atmospheres are termed: 1) tropical, 2) mid-latitude summer, 3) mid-latitude winter, 4) subartic summer, 5) subartic winter, and 6) 1976 U.S. standard atmosphere.

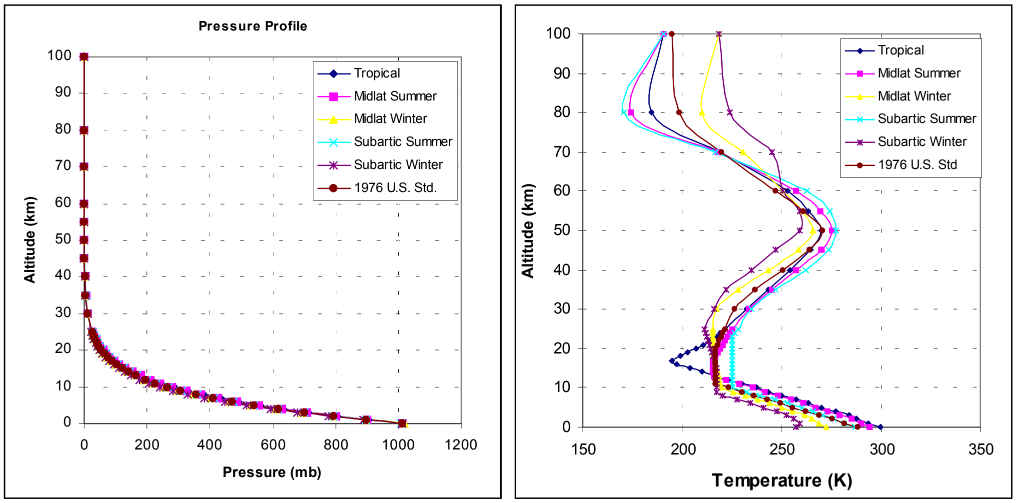


Figure 7 Atmospheric profiles for vertical pressure and temperature.

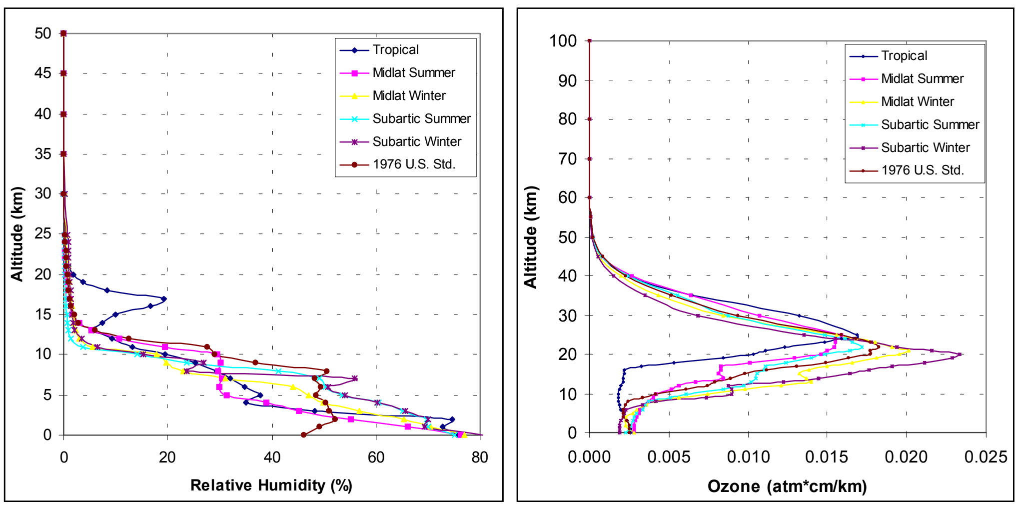


Figure 8 Atmospheric profiles for vertical relative humidity and ozone.

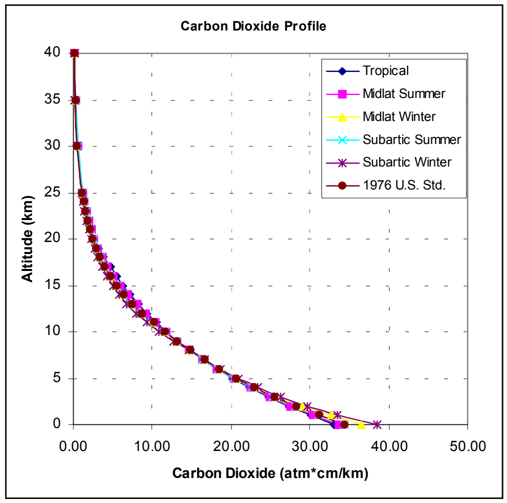


Figure 9 Atmospheric profile for vertical carbon dioxide.

The factorial design of the three major parameters resulted in $6^3 = 216$ observations. An illustration of this can be seen in Figure 10 and Figure 11. No repeats were measured because the model is a physical model and the results are not random variables.

The output from the MODTRAN model is the simulated observed radiance from a sensor at an altitude of 100 km. The bandpass for the runs was between 7.34 and 13.57 µm (737 to 1362 cm^{-1} at 5 cm^{-1} intervals). All of the observations were made assuming a nadir sensor-target geometry.

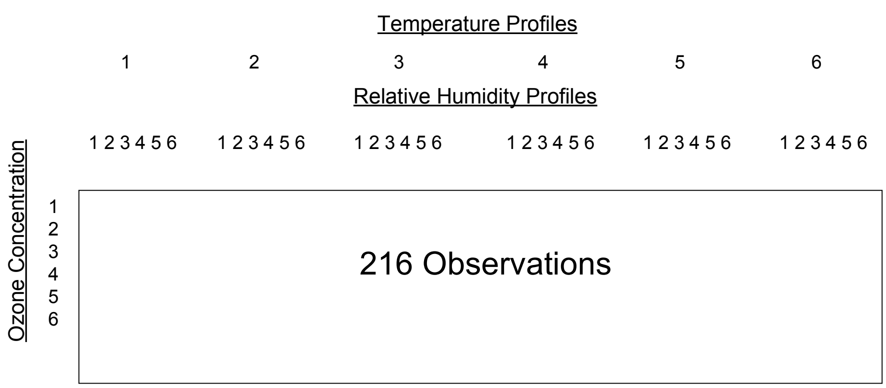


Figure 10 Factorial design of major input variables.

MODTRAN Run					
1	P(z)	$T(z)_1$	$RH(z)_1$	$03(z)_1$	CO2(z)
2	P(z)	$T(z)_1$	$RH(z)_1$	$03(z)_2$	CO2(z)
⋮	⋮	⋮	⋮	⋮	⋮
6	P(z)	$T(z)_1$	$RH(z)_1$	$03(z)_6$	CO2(z)
7	P(z)	$T(z)_1$	$RH(z)_2$	$03(z)_1$	CO2(z)
⋮	⋮	⋮	⋮	⋮	⋮
36	P(z)	$T(z)_1$	$RH(z)_6$	$03(z)_6$	CO2(z)
37	P(z)	$T(z)_2$	$RH(z)_1$	$03(z)_1$	CO2(z)
⋮	⋮	⋮	⋮	⋮	⋮
216	P(z)	$T(z)_6$	$RH(z)_6$	$03(z)_6$	CO2(z)

Figure 11 Illustration showing make up of each MODTRAN run.

4 Results

The overall experimental design can be seen in Figure 12. Here we have the 216 atmospheric profiles as inputs to MODTRAN which generates 216 observed spectral radiances. Additionally, MODTRAN generates 216 transmissions and 216 upwelled (or path) radiances. In reality, MODTRAN generates much more than this. However, for this experiment, we will only concentrate our efforts on the above mentioned parameters. We then have to train the model with the pair of parameters of interest, in order to generate our regression matrix, β_{cc}. So if we wish to estimate a temperature profile from an observed radiance, for example, we would train with 216 temperature profiles and their corresponding 216 observed radiances. As illustrated in Figure 12, we then feed an observed radiance into the CCR inverse model to generate an estimated atmospheric profile or spectra (depending on the training). Lastly we compare our estimates to the actual profile or spectra.

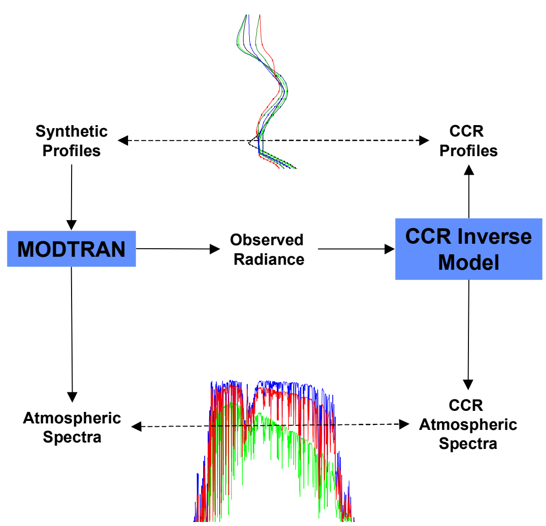


Figure 12 Experimental design using 216 atmospheric profiles.

4.1 Estimating Atmospheric Temperature Profiles

The first experiment was to try and estimate a temperature profile from an observed radiance. After training, we end up with 36 canonical correlations. The first 8 of which can be seen in Table 1. In developing the CCR model, we only keep the significant correlations whose running sum is 85% of the total variability or correlation in the data. So for temperature, we only kept the first 4 canonical correlations which explained 88% of the variability. It would make sense to add the fifth correlation, bringing the total variability to about 100%, however, in this experiment we wanted to choose something slightly less than optimal for generating estimates. Certainly we would not choose more than 5 correlations for temperature estimation. As can be seen in Table 1, we do not gain any additional information by including more than 5 correlations. The correlations for all the estimated parameters can also be seen graphically in Figure 13.

Table 1 The first 8 canonical correlation's and their percentage of total variability for temperature (T), transmission (τ), and path radiance (L_u).

Correlation (T)	0.99986	0.99944	0.93067	0.85834	0.50947	2.9E-16	2.9E-16	2.6E-16
Total Variability	0.23	0.47	0.68	0.88	1	1	1	1
Correlation (τ)	0.99994	0.99922	0.99545	0.99007	0.97106	0.88189	0.84875	0.71695
Total Variability	0.14	0.27	0.40	0.54	0.67	0.79	0.90	1
Correlation (L_u)	0.99997	0.99887	0.9941	0.98559	0.9648	0.89849	0.81156	0.31977
Total Variability	0.14	0.29	0.43	0.57	0.71	0.84	0.95	1

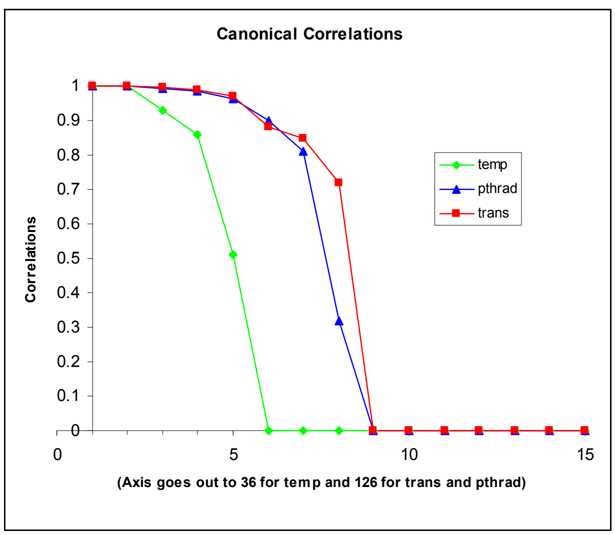


Figure 13 Canonical correlations for temperature, transmission, and path radiance.

The results for the estimation of temperature can be seen in Figure 14. Here we trained the model, as stated above, kept 4 canonical correlations, then use the regression matrix to estimate Y from X, where Y was 216 temperature profiles and X was 216 observed radiances. Figure 14 shows the result for observation #80 [1row x 36col]. When compared to the original observation

#80, we see (in the residual plot) that we can estimate temperature to with in ±0.5 K. This can, however, be misleading. This result is valid for observation #80, not the entire ensemble of observations. Therefore, we need to look at the error for all the observations. This can be achieved and presented as a histogram, as can be seen in Figure 15. Here we see that most of the errors are with in ±1.0 K. We also notice that the errors tend to be normally distributed.

a) b)

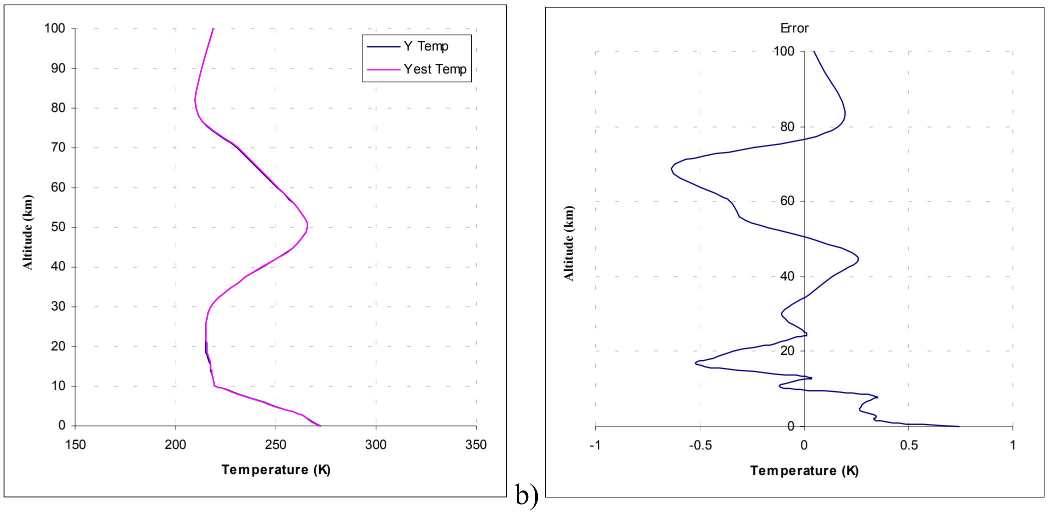


Figure 14 a) Results of estimating atmospheric temperature profile #80. b) Residual temperature plot for profile #80.

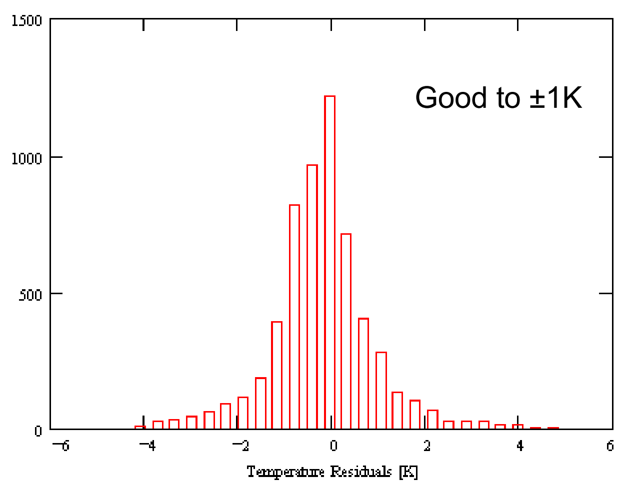


Figure 15 Histogram of residuals for temperature estimation.

4.2 Estimating Atmospheric Transmission

We now estimate atmospheric transmission. This is actually an *output* parameter from MODTRAN, along with the observed radiance, rather than an input parameter. After training, we end up with 126 canonical correlations, which is equivalent to the number of bands in the region of interest. Again, the first 8 can be seen in Table 1. This time we end up keeping 7 correlations, unlike the 4 we kept for temperature estimation. The results of the transmission estimation can be seen in Figure 16 along with the residual plot. The y-axis of the plot is in percent transmission, which can range from 0-100%. The histogram of all the residuals can be seen in Figure 17. Here we see that most of the errors fall with in ±1.0 percent or 0.01. Again, we see that the errors tend to be normally distributed.

a) b)

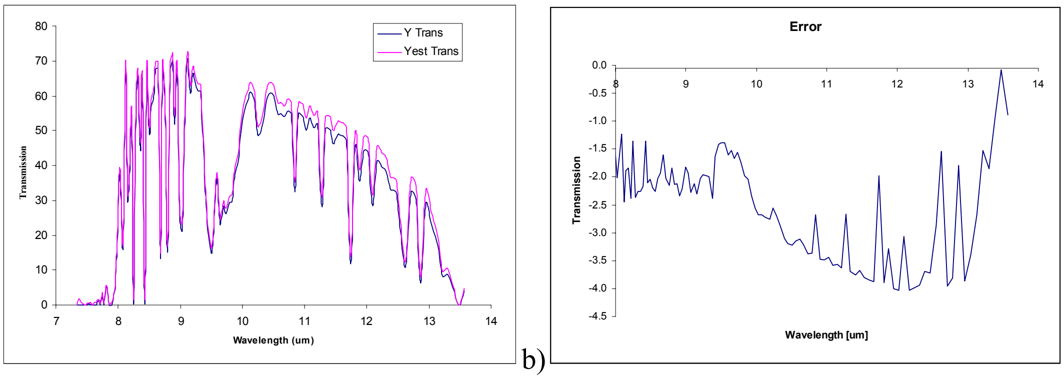


Figure 16 a) Results of estimating atmospheric transmission profile #2. b) Residual transmisson plot for profile #2.

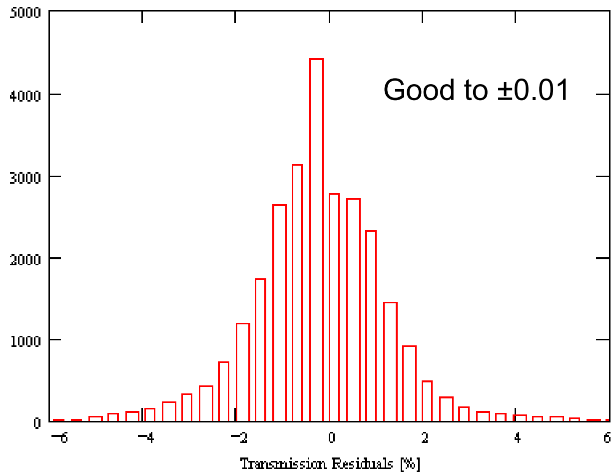
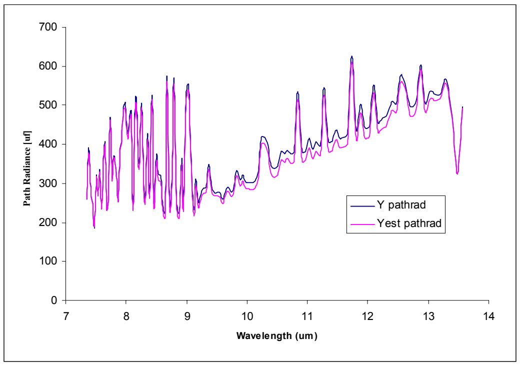


Figure 17 Histogram of residuals for transmission estimation.

4.3 Estimating Atmospheric Path Radiance

Finally, we use CCRA to estimate path radiance. Path radiance is sometimes referred to as *upwelled radiance*. It is the radiance that emanates from the surround and falls into the path or line of sight of a sensor. This radiance or energy does not originate from the target. It is merely the energy from outside the sensors line of sight that makes it onto the sensors focal plane along with the reflected targets energy. Along with atmospheric transmission, path radiance is also an output parameter from MODTRAN. Upon training we again end up with 126 canonical correlations. The first 8 of these can be seen in Table 1 and graphically in Figure 13. Again we keep 7 correlations which explain 95% of the variability. The results of the estimation for profile

#2 can be see in Figure 18 along with a residual plot. The histogram of residuals (Figure 19) shows that most of the errors fall with in ±15.0 µf, where µf is an abbreviation for a microflick. A microflick is equivalent to 1 [µW/cm^{-2} sr µm], which is the traditional unit for spectral radiance.



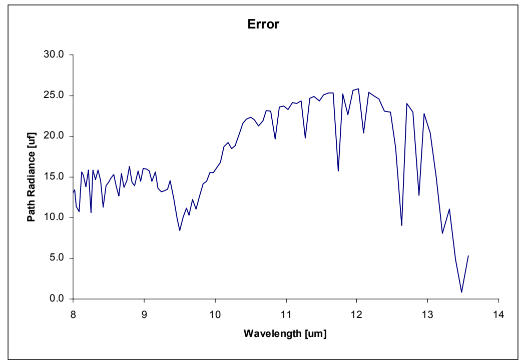


Figure 18 a) Results of estimating atmospheric path radiance profile #2. b) Residual path radiance plot for profile #2.

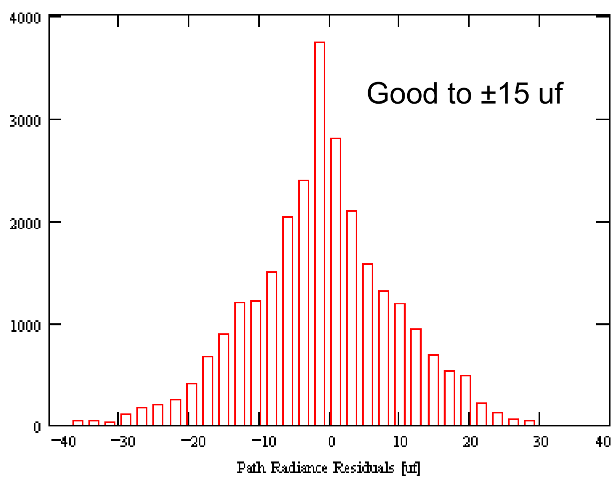


Figure 19 Histogram of residuals for path radiance estimation.

5 Conclusions

The above results clearly show that canonical correlation analysis used in a regression scheme can adequately be used to predict atmospheric parameters such as temperature profiles, spectral transmissions, and spectral path radiances. The original goal was to develop a function that was invertible and that could faithfully represent the nature of the physics based model. In this report, the physics based model was MODTRAN. With that said, we have successfully built an inverse model by using CCA as a rank-reduced multivariate regression to capture all the relevant physics contain in the physics based model. The extension of this approach is to use real data from radiosonde balloons as input to the physics based model. This would then generate observed spectra to use as training, in conjunction with the radiosonde data. Again, we consistently put our faith in the MODTRAN model, only because it has been validated and tested throughout the community for the last 30 years. Lastly, we would take imagery, which contains real observed radiances, and use it in our inverse model to predict our atmospheric parameters of interest. Hernandez-Baquero used this approach with radiance data from a handful of multi and hyperspectral imaging sensors to estimate atmospheric parameters as well as emissivity and surface temperatures.

6 References

Hernandez-Baquero, E.D., “Characterization of the Earth’s Surface and Atmosphere from Multispectral and Hyperspectral Thermal Imagery,” Ph.D. Thesis, Rochester Institute of Technology, Center for Imaging Science, June (2000).