

# Analysis of the Arecibo World Day Data, October 1985 to May 1995

Michael P. Sulzer

June 1995

DRAFT

## 1 Introduction

The purpose of this document is to describe the reanalysis of the Arecibo world day data from October, 1985 to May, 1995. It contains some introductory material which should be ignored by those familiar with the subject of incoherent scatter data analysis, but it also contains some new material. The overall approach is to process the data before least squares fitting to obtain independent spectra at each range. This is more complex than the simple triangular weighting that has been used in the past to correct the autocorrelation functions. It involves deconvolving the profiles of lag versus range. It is as good as the full profile fitting technique (in the case when range profiles of the physical parameters are not available). The use of these correction techniques can make as much as a 200 degree difference in the electron temperature during the daytime where the electron temperature profile peaks in the lower F region. Differences in ion temperature are smaller. The technique used here is much simpler and faster than the full profile technique; it is shown it is as good.

Constrained fitting is used at the higher ranges where three ions must be allowed free. Error bars are lowered by as much as a factor of three. The constraint technique is derived and some differences from Erickson and Swartz (1994) are discussed.

The L+M fitting described in Numerical Recipes has been used throughout the project. It is very important that the best possible algorithm be used in difficult cases; otherwise, limitations in the results are perceived as inherent rather than as resulting from an inferior technique.

When the project began, the theoretical ACFs were generated using code written by Swartz (1978a and b). The last two passes through the data have used new code for generating the spectra and derivatives, in which the special case at zero frequency has been eliminated, making the code simpler, and the making of the theoretical and practical spectra (based on a finite length ACF) has been divided into two steps, making the process easier to understand. The new code is somewhat less efficient than the old, but this is not a problem given the current, and especially, future speed of inexpensive computers.

The velocities are essentially unchanged in the new analysis. In one case (January, 1986), the transmitter chirp had been incorrectly removed, but little else has changed. Also the temperatures near the peak of the F region, where  $O^+$  predominates, are little changed. However, at the higher altitudes the ion fractions and temperatures are quite different. Light ion fractions are significantly higher at solar minimum than has been commonly believed, and the old technique for handling the noise baseline could not give accurate results in such cases. The new technique is described below.

## 2 Description of the Experiment

The radar technique used for the world day experiment at Arecibo beginning in October, 1985 and continuing until the present is described in Sulzer (1986). Seven frequencies are transmitted by phase modulating the carrier with a repeating binary waveform. The frequencies are spaced by about 35 KHz, and a 250 KHz bandwidth, containing the returns from the seven frequencies, is analyzed. Figure 1 shows sets of seven spectra, from the highest four ranges of the experiment, before the removal of the noise baseline. These data are from a recent (solar minimum) experiment with low electron densities; the fraction of  $H^+$  increases rapidly with height. These spectra might appear unfamiliar to those familiar with only the nearly rectangular shape of an  $O^+$  spectrum from the middle of the F region, but it is our intention to show more difficult cases, since the easy cases do not require the techniques described here. The data are from 3:16 on April 25, 1995. Data from lower heights are much stronger and show little effect of the filter shape, and these data are not shown since there is less difficulty in preparing them for fitting. The actual filter shape is very flat but does not fall off quickly enough to prevent the folding in of a significant amount of noise from outside the clear bandwidth. This is the explanation for the rising baseline near the edges of the bandpass. The method for removing the noise baseline is discussed below; it is more difficult than it might be because no noise samples were taken along with the data. Thus the amplitude of the noise spectrum is not known, although the shape is known because it changes little with time and can be determined from times possibly several hours away from the data, when the transmitter was not operating, but the data-taking program was, typically with the receiver connected to a 300 degree resistor rather than the antenna.

Figure 2 shows the same spectra after the removal of the noise baseline. The variations in amplitude among the seven spectra from a single range are the result of the imperfect binary phase code used to generate the frequencies and of imperfections in the transmitter, resulting in the asymmetry about 430 MHz. The larger narrower spectra are from the lower heights. As the height increases, the amount of  $O^+$  decreases quickly, leaving the wider  $H^+$  spectrum. The shape of this latter spectrum is not completely resolved because the spectra are too close together. That any shape at all is apparent is due to the low temperature of the solar minimum night. After sunrise, the shape becomes completely flat, and we rely entirely upon the level of the spectrum to determine the fraction of  $H^+$ . Getting a temperature measurement without a significant fraction of  $O^+$  is not possible with this data.

Figure 3 shows the spectra after the seven frequencies have been combined into a single right half spectrum shifted to zero frequency. This is the spectrum used in the modeling. It is the Fourier transform of the truncated autocorrelation function, where the lags all have uniform weighting. A careful examination of figure 3 shows that the spectra from the lower heights cross the spectrum from 683 km near the middle of the frequency range, but that they do not do so in figure 2. The oxygen component is sufficiently narrow so that the triangular weighting applied to the autocorrelation function by the radar technique is significant, and thus its removal causes a spectrum containing a significant amount of  $O^+$  to narrow. Although the non-linear least squares fitting program does not care what form the data are in (as long as all effects are accounted for), we have found that the spectrum based on the uniformly weighted truncated ACF is best for visual inspection of the results because it shows the most information to the human observer.

## 3 Modeling Incoherent Scatter Data

The object of modeling experimental data is to reduce many numbers to a few parameters which describe the physical situation to the best degree possible. When one

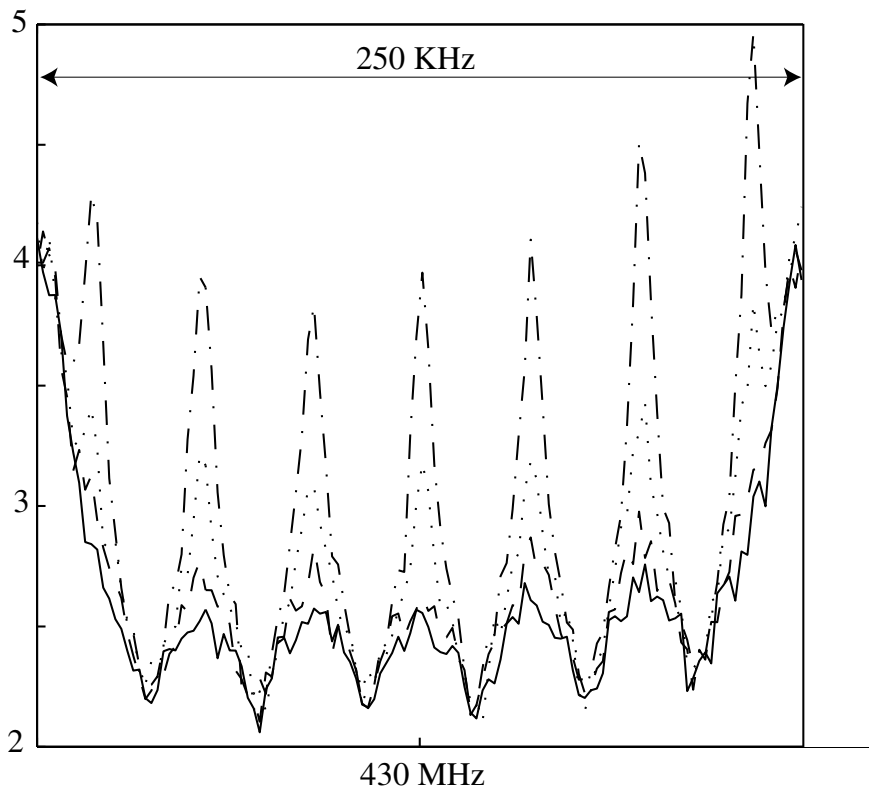


Figure 1: Spectra from the four highest ranges, 568 (dash-dot), 606 (dotted), 645 (dash), and 683 (solid) km. The noise baseline has not been removed; its two dominant features are raising the entire level of the spectrum well above the level of the ionospheric spectra, and causing the level to rise at the edges of the spectrum.

attempts to design a technique for associating a set of parameters with a data set, one finds three important characteristics of the data. First is what might be called the physics of the situation: there is a function of the parameters which can be compared to the data, not the actual data, but rather an idealization which is what the data would be like without certain complicating factors which occur in any real experiment. For incoherent scatter, this is the function which relates parameters such as the electron and ion temperatures and ion masses to the theoretical spectrum as a function of altitude. The second characteristic is that which relates this function to actual data; these are limitations in the measuring technique and so on. For example, in radar work one finds that the length of the pulse must be accounted for; some effects would go to zero only if the time duration of the pulse were infinite, while others would vanish only if the time duration were zero. Thus there must always be some effects, and in general they are very significant. The third is random errors in the data values; these place a limit on the accuracy of the parameters associated with the data. Of course a poor model can result in larger errors than those given by the noise, but if so, then the modeling process has not been fully successful, and it might be necessary to improve either the analysis, the experimental technique, or both. The goal for the model is that it be as good as or better than required to allow the statistical errors to dominate. There are techniques for determining if this is the case, and it is very important that they be carefully applied. Otherwise it is not possible to know the accuracy of the

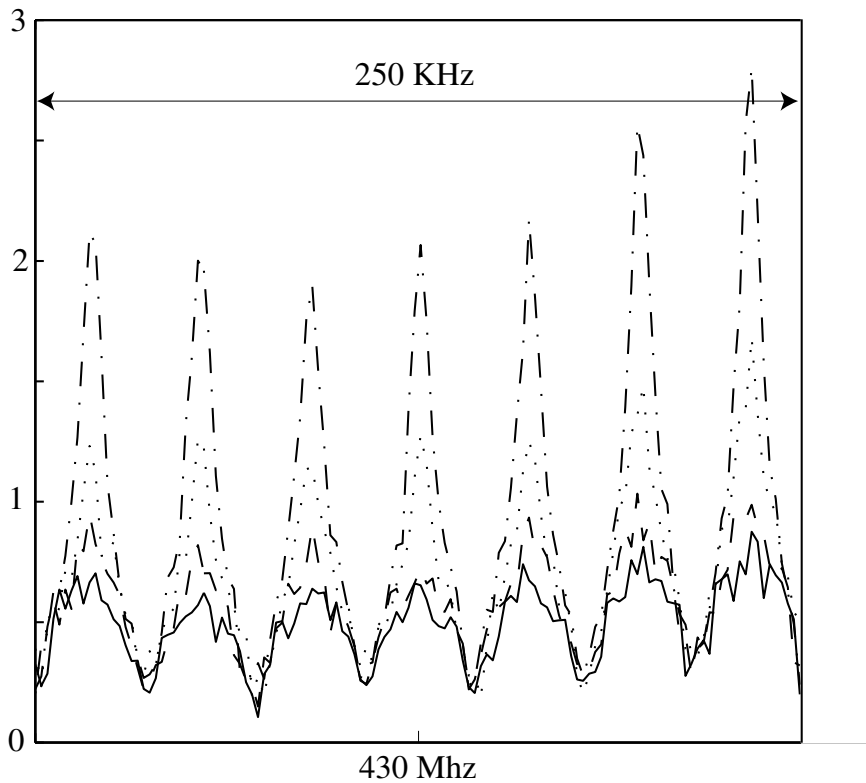


Figure 2: The same data as Figure 1 but with the noise baseline level removed. The spectra are still (apparently) above the baseline; this is caused by the presence of  $H^+$ . The wide spectra of  $H^+$  are incompletely separated and thus cause the entire set of spectra to appear raised above the zero level.

parameters or how to do better in the future.

We use non-linear least squares fitting for generating the parameters which belong to a data set. (See Swartz (1978a and b) and references therein.) The data are compared to the model by subtracting, squaring, weighting by the inverse square of statistical errors, and summing. The parameters which give the minimum sum are accepted as the answers. A strict interpretation of this process would require that the data used are the numbers that are read from the data tape or other storage medium and that the model must contain both the physics and the other non-ideal characteristics. This view is not very realistic; and we find that it is much better to take a more general view of what the data are. For example, the radar data under consideration here represent the sum of spectral estimates from many radar pulses. These spectral estimates are special in that they are computed using a technique which is intended to allow the computation of undistorted autocorrelation functions (ACFs). One might argue that these ACFs are actually the data, but a purist could argue that the Fourier transform required to recover the ACFs from the spectra takes one further from the original process and therefore we should use the spectra as the data. However, a further look at the on-line processing shows more trouble for one looking for the 'actual' data because the spectra accumulated on line contain twice the number of points as the ACFs (or the spectra written to tape). This is because the computational technique is designed for the maximum speed and results in a complex spectrum in which the

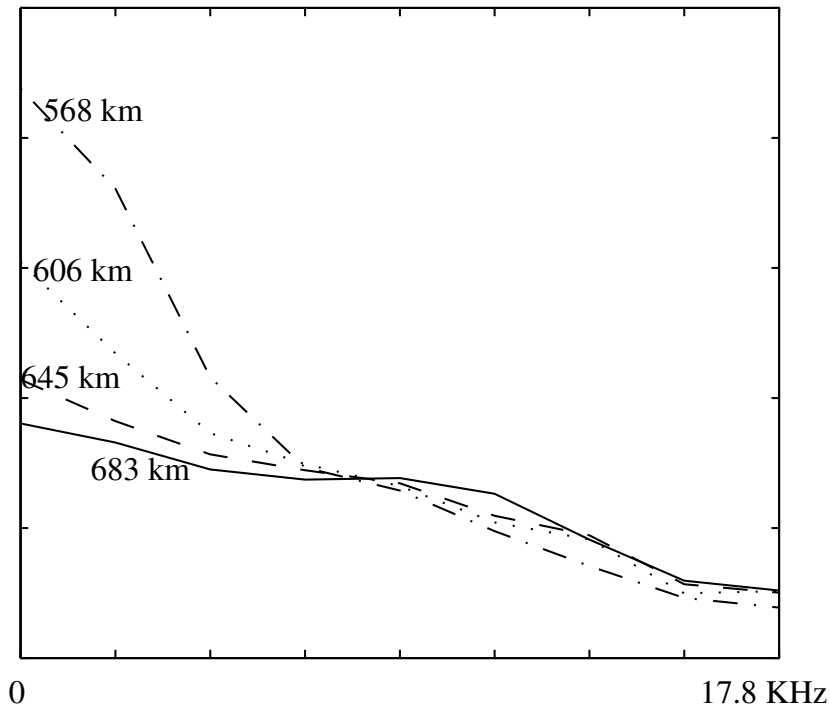


Figure 3: The same data as the two previous figures, but after the seven spectra have been combined into one. This is the form of the spectrum used in the fitting process.

wanted spectra are combined with useless numbers. To extract the information, the complex spectrum is Fourier transformed, and certain lags of the resulting ACF are selected, and the rest discarded. One could have written the ACF to tape, but it is more convenient to monitor the spectrum, and so the ACFs are transformed back to a real spectrum. We argue that neither the spectrum nor the ACF is fundamental, and that the concept of constructing a model to fit to the actual experimental data (as if they were written down by someone in a white coat) is too naive to apply to radar data, or for that matter to most other scientific data. We also argue that one can apply any reversible linear transformation to the data before the fitting process. To justify this one must show that the parameters obtained before the transformation also would be obtained after any such transformation. We will show below that it is true for some transformations, namely convolutions. For the general linear transformation one can see that if our contention is not true, then there are some serious problems with data analysis. If some transformations give better results than others, surely it is unlikely that our data just happen to be in the best form. We would have to find which form is best; since this is not done in general, one could argue that all data should be reanalyzed after the correct procedures are developed.

The reason one wants to transform the data is to obtain a representation that is most convenient. For example, one can remove many of the distortions in the data (characteristics of the second kind defined above) before fitting. This simplifies the model and allows the data set (in this case) to be divided into subsets which can be fit independently. One can also choose a representation in which the effects of varying the various parameters are most obvious to the eye. This can be a big help when the fitting process is developed and monitored.

## 4 Corrections to the data before fitting

We now discuss briefly the kinds of distortions the data contain and the choices made for removing them. A fuller discussion of some of these is given in the next few sections. First is excess power at zero frequency (often referred to as DC), caused by systematic offsets in baseband mixers, operational amplifiers, and A/D converters. This is a problem because so called baseband sampling is used. A pair of real signals represents a complex signal, and a DC offset in either of the signal paths appears as excess power in the middle of the spectrum. In principle the offset can be removed by keeping track of the average values of the samples at the various ranges, but actually this does not quite work for unknown reasons. Since we have seven spectra and only one of them has the DC problem, we replace the bad point with an estimate based of the other six. We will not discuss the algorithm in detail.

The next problem is additive noise, which appears as an additive signal at each frequency in the spectrum. (Here we are referring to the actual level or expected value of the noise signal, not the fluctuations in it, which of course average to zero.) Over most of the spectrum the level of the additive signal is proportional to the response of the filter at that frequency, but near the edges of the spectrum there is an additional component resulting from aliasing. Since the filter does not have perfectly steep sides, noise from outside the clear bandwidth set by the sampling rate folds into the desired spectrum, complicating the task of removing the noise. No associated noise samples were collected with this data so that calculations for as many ranges as possible of data could be performed. This choice was made because the noise removal is not a problem for a spectrum resulting from a medium containing only oxygen and a small fraction of Hydrogen ions. In this case the only effect of the noise is on the zero lag of the ACF made by combining the seven spectra, and the zero lag can be discarded. (The shape of the specially designed baseband filter is such that the effect of noise on the other lags of the ACF is very small.) However, when light ions at much greater than expected levels were encountered in some of the data, it was necessary to devise a method for removing the noise level. This is described in a later section.

The most difficult and sometimes misleading problem is distortion introduced by the finite length of the radar pulse. The problem is most easily understood in the time domain: if the medium is represented by a matrix of ideal ACFs, one for each range, then the actual ACFs are related to the ideal ones by a convolution in the range direction, where there is a different convolving function for each delay time  $\tau$ . The most general case is more complicated than this, but the good properties of the baseband filters allow this simplified treatment. The most straightforward general solution for this problem is to build the range convolution into the model and fit for all ranges at once. This is complicated and computationally intensive. We show in a later section that the range convolution can be removed just as well before the fitting, leaving spectra which are independent in height which can be fit separately.

The last problem is really somewhat different in nature from the others, but we include it here because it can be solved in a similar way to the others with a technique included in the analysis. The problem is that when the signal to noise ratio is low, a model with sufficient degrees of freedom to allow good fits results in parameter values with very large random errors. The solution is to restrict the range variation of parameters known to change slowly in range; the method is described in a later section.

## 5 Removing the effects of additive noise

Since no noise samples are associated with each set of spectra, the simple procedure in which the noise spectrum is subtracted from the spectra containing data is not possible. However, the shape of the filter response to noise is known; usually it can be derived

from several records in a data set where no data are present. This usually occurs because the transmitter was not functioning for a short time. In the rare cases where such records are not available, a standard measurement of the filter shape can be used. The baseband filters use operational amplifiers and stable passive components, and the responses have not changed significantly over a ten year period. However, certain special experiments with complicated setups can modify the shape of the passband and care must be taken to find the correct response in these cases.

With the shape of the response known, there is only one unknown, the amplitude of the noise. To find this, we take advantage of certain characteristics of the filter response, in particular, the fact that it has very small ripple and the points of zero derivative in the response are located at the centers of the seven received spectra. Other filters in the system, such as at the intermediate frequency level, also make some contribution to the shape, but the dominant shape is due to the aliasing of noise from outside the clear bandpass set by the sampling rate.

The noise level is found by means of a linear least squares fit to a function derived from the spectrum containing the seven frequency return from the highest altitude where the signal is weak and so the filter shape is least obscured by the incoherent scatter. The fit resolves the fitted function into a sum two components; each component has a fixed shape. Thus the only degrees of freedom are the two coefficients which give the amplitude of each component. The first component is derived from the measured filter spectrum, and consists of seven points; each point is the sum of five points from the measured spectrum. The seven points are centered halfway between the transmitted frequencies so that when oxygen is the dominant ion, as little as possible incoherent scatter is included in the sum. This component has a distinctive shape resulting primarily from the noise folded in from outside the nominal bandpass. The other component is derived from the spectra measured at the peak of the F region, or it can also be obtained from the spectrum of the transmitter sample. It consists of seven numbers which are the relative powers of the seven transmitted frequencies. It has a distinctive shape because the seven frequencies are transmitted with unequal powers. Since the shapes of the two components are very different, the relative levels of the two can be determined very accurately by the fit. The coefficient of the first component is multiplied into the filter spectrum and the result is subtracted from all of the spectra.

## 6 The effects of range smearing and how they can be removed

The purpose of this section is to prove the following assertion: the non-ideal effects introduced into the radar data by the finite length of the pulse, called range smearing, are equivalent to passing a signal through a filter, and they can be reduced, or completely removed under some conditions, by passing the signal through another filter. For the purposes of assessing range smearing the plasma is best represented by a set of autocorrelation functions, one for each range. For convenience this can be thought of as a set of continuous functions in range  $\rho_{\tau}^i(r)$ , where  $\tau$  is delay, that is the lag value of the acf,  $r$  is the range, and  $i$  indicates that this is the ideal function of the medium without smearing effects introduced by the measurement. It is most convenient to think of  $\tau$  as varying in the horizontal direction and  $r$  in the vertical direction. As will be proved below, this ideal function is converted into the measured (data) function,  $\rho_{\tau}^d(r)$ , by a set of convolutions in the vertical direction. That is, for each value of  $\tau$ , there is a function  $h_{\tau}(r)$  which gives the measured function for  $\tau$  when convolved into the ideal function for that fixed value  $\tau$ :  $\rho_{\tau}^d(r) = h_{\tau}(r) * \rho_{\tau}^i(r)$ . This does not describe the whole problem, since there is a convolution in the horizontal direction as well, but this effect can be made negligible by using good filters, or filters that are

wider than strictly necessary and performing the extra computation necessary for the wider bandwidth. The filter used for the data described in this paper is sufficiently good so that only the convolution in the vertical direction is significant.

An analogy is useful here. For each value of  $\tau$ , the convolution in the vertical direction is like passing a signal through a filter. An even more specific analogy is appropriate. When a signal passes through an analog long distance phone line, the high frequency response is decreased. This effect is one of filtering, that is, convolution by the impulse response of the phone line. An importance difference is that the phone signal has indefinite length while the radar signal is computed only over a finite distance, and so we must be careful to allow for end effects by computing the radar signal over a larger range than we actually need it, or to ranges where it is negligible. To continue the analogy, the phone signal might well be unintelligible after passing through the line, but it can be much improved by equalization, that is, by passing it through a filter that undoes the loss in the high frequencies. Neglecting other effects such as non-linearities in amplifiers, we can do a very good job of fixing the phone signal, limited by noise in the system, which of course is increased by the filter along with the signal. If the loss of high frequencies is too great, the noise will be too large compared to the signal and it will be unintelligible even after equalization.

The radar pulse acts as a filter to take out the higher spatial frequencies in the scattered signal. The signal can be filtered to increase these frequencies, and thus  $\rho_\tau^d(r)$  can be made closer to  $\rho_\tau^i(r)$ . The filter shape introduced by the square radar pulse is not very good for this purpose; the Fourier transform of a square pulse has a shape given by  $(\sin \omega)/\omega$  which has zeros in its response. Amplification of frequencies near the zeros will result in very large increases in the noise level, and thus there will be a large increase in statistical errors if we need to correct spatial frequencies that are near the zeros. If only a small increase in resolution is required, only frequencies below the first zero need be affected, and there will be a small increase in noise. However in cases where a significant increase in resolution is required, we need to find a way to get the spatial frequencies that are missing.

For the data considered in this paper, the first case applies, that is, we are making only small corrections in the data, and furthermore, the convolution effects are only significant below the peak of the F region. Above the peak, the range variations of  $\rho_\tau^d(r)$  are very close to those of  $\rho_\tau^i(r)$  with the  $308\mu$  sec pulse used in the experiment, while in general there are significant effects below the peak. Significance is measured with respect to the size of the statistical errors: when the correct errors are used to set the  $\sigma$ s in the  $\chi^2$  equation the resulting value of  $\chi^2$  falls within a range of values if the model is good. The effect of range smearing is to raise  $\chi^2$  out of the permissible range. Correction techniques can in principle fix the deficiencies in the model, but the data considered here cannot be completely corrected because  $\rho_\tau^d(r)$  is sampled with too large a range interval. As a result, the errors associated with these data must be increased above the size of the statistical errors to allow for the systematic effects.

The large sampling interval in  $\rho_\tau^d(r)$  means that some of the information necessary for complete correction is missing. It might appear that the full profile technique would be best for handling this missing information, and this would be true if we had physical models for the parameters such as the temperatures and densities. However, we do not have such models, and any model which we construct for these parameters is just a form of interpolation. However, there is no reason why interpolation of the physical parameters is any better than interpolation of the lag profiles themselves, since they can be assumed to vary smoothly between the samples just as well as the physical parameters. Interpolation is incorporated into the filtering process by using an iterative technique in which the filter inverse is never used. This technique also has the advantage of avoiding any problems with infinities resulting from zeros in the filter response which could blow up in the inverse. The technique uses  $\rho_\tau^d(r)$  as the first approximation to  $\rho_\tau^i(r)$ . The sampling rate of this approximation is increased by cubic

spline interpolation and it is passed through the filter ( $h_\tau(r)$ ). The result is scaled, shifted in range and subtracted from the first approximation. The differences are used to make the second approximation, and the process is repeated. Two passes are sufficient for the process to converge because the required corrections are small. Of course this does not work on the very steep bottom side gradient of the nighttime F region, but for that we need more information. The sampling interval must be decreased by several times, and additional spatial information must be supplied by using more than one pulse length. For this reason, the gradient of the power profile is measured and all data from regions with a gradient above a selected threshold are rejected.

Now let us proceed with the proof that  $\rho_\tau^d(r)$  and  $\rho_\tau^i(r)$  are related by a set of convolutions. Consider an ionosphere that consists of a single layer so thin that it can be considered an impulse. If we transmit a radar pulse of length  $T^t$ , we receive a signal  $s_{ii}(r)$ , where  $r$  is the range variable and  $ii$  indicates the ionospheric impulse. The autocorrelation versus range ( $\rho_\tau^m(r)$ ) for delay  $\tau$  is the expected value of  $s_{ii}(r)s_{ii}(r - \tau)^*$ ; that is, a pulse of length  $T^t - \tau$ . This is true no matter where the ionospheric impulse is located, and thus so far this is consistent with representing  $\rho_\tau^d(r)$  as the convolution of the medium with some function which depends on  $\tau$ ; that is, we have satisfied the shifting requirement of convolution. To finish the proof, we must show linearity, that is, that the  $\rho_\tau^d(r)$  resulting from a medium with scattering from a range of heights is the sum of the functions obtained by considering each range separately. The truth of this assertion is dependent upon the properties of the medium. For incoherent scatter, the response at one height is independent from that at another, and so the autocorrelation functions add. This completes the proof.

## 7 Comparison to the full profile technique

When correcting the range smearing described above we will make use of this: we can convolve the radar signal with some function, meeting some restrictions, without losing any of the information that is contained in the signal. This allows a two step approach in the data analysis in which the range convolution is removed first and then the different ranges are fit independently for the geophysical parameters using non-linear least squares fitting. An alternative approach would involve fitting all heights and lags at once, including the convolving effects of the long pulse in the model used in the fitting. The full profile method is generally recognized as extracting all the useful information from the data. However, it is complicated, and therefore we must determine if it is necessary. We will show that the data can be convolved by a function meeting certain restrictions without changing the results obtained by fitting. This result can easily be extended to a set of convolutions on subsets of the data. Then if the convolutions produce independent subsets, each subset can be fit separately.

Consider a signal (with additive noise) which depends on several parameters; We would like to find the values of the parameters which best describe the signal, using non-linear least squares fitting. Suppose we convolve the signal with another function; we will show that exactly the same parameters (to within computational limits) result from fitting either the original function or the convolved function. Let the signal be a function of time, and the convolution be realized by passing the signal through a filter. The signal  $y(t)$  approaches zero at its ends and we have plenty of samples before and after so that there is no problem with end effects after the convolution. Also the samples of the signal are closely spaced so that it is well-represented; that is, the signal  $y(i\Delta t)$ , where  $\Delta t$  is the sampling interval, is not significantly aliased.

First, let us look at a qualitative proof that shows that the filter causes no loss of information. Take the Fourier transform of the signal; at each frequency we have both signal and noise. Now take the Fourier transform of the signal after passing through the filter. The two transforms are different, but the convolution theorem tells us that the two are related by multiplication by some function. That is, if  $z$  is  $y$

after passing through the filter, and upper case represents the Fourier transform, then  $Z(f) = H(f)Y(f)$ . Thus, the amplitude and phase can be altered at each frequency. Note that the signal and the additive noise are both affected by the same amount and so the signal to noise ratio at each frequency is not affected. If the signal to noise ratio has not changed at any frequency, then we have not lost any information. We assume that the attenuation of the filter at all frequencies is small enough that the finite accuracy of the computations does not add significant noise. Thus we must avoid filters that have zeros in their responses.

Now we show that the fitting gives the same information before and after the convolution (filter). The fitting process consists of computing the  $\chi^2$  merit function for a particular set of parameters, and varying the parameters until  $\chi^2$  is minimized. We will show that  $y$  and  $z$  have the same  $\chi^2$  function and thus the minimum is at the same parameter values. For the purpose of the proof it is convenient to fit to  $Y$  and  $Z$ . Fitting to the Fourier transform makes no difference; for example one can fit either to the ACF or its transform, the spectrum, with identical results, assuming the correct  $\sigma$ s are used. We assume that all functions are real, but the proof can be easily extended to complex functions. For  $Y(f)$  the merit function is

$$\chi_Y^2 = \sum_i \frac{(Y_i - Y_{mi}(\mathbf{a}))^2}{\sigma_{Y_i}^2},$$

where  $Y_i$  is  $Y(i\Delta f)$ ,  $Y_m(\mathbf{a})$  is the model for the vector of parameters  $\mathbf{a}$ ,  $\sigma$  is the error associated with each data point  $Y_i$ , and the sum is over all data points.

At this point we can easily prove the assertion made just above that the same results are obtained with either the data or its Fourier transform, at least for the case where all of the  $\sigma$ s are the same. Referring to the last equation, one sees, by applying the linearity property, that the difference between the data and the model is  $y_i - y_{mi}$  in the time domain and that this is the Fourier transform of the quantity in the equation. Taking the constant  $\sigma$ s outside the sum, it is apparent that the sum of squares is the same in either domain, since the total power is the same in either domain (assuming proper normalization) by Parseval's theorem. Thus  $\chi^2$  is the same function in either domain.

For  $Z(f)$  the merit function is

$$\chi_Z^2 = \sum_i \frac{(Z_i - Z_{mi}(\mathbf{a}))^2}{\sigma_{Z_i}^2}.$$

Using the relationship  $Z(f) = H(f)Y(f)$ , we have

$$\chi_Z^2 = \sum_i \frac{(Y_i - Y_{mi}(\mathbf{a}))^2 / H_i^2}{\sigma_{Z_i}^2}.$$

If we let  $\sigma_{Z_i} = \sigma_{Y_i} / H_i$ , restricting  $H_i$  to have an amplitude always significantly greater than zero, we have

$$\chi_Z^2 = \sum_i \frac{(Y_i - Y_{mi}(\mathbf{a}))^2}{\sigma_{Y_i}^2},$$

and the equality is established.

The radar application is a bit more complicated: the data are a two dimensional function, range and delay. We do a set of convolutions in the range direction, on for each delay value  $\tau$ . A simple extension of the proof above shows that we still have the same information. We can fit the entire data set at once if we want (as we are required to do before the deconvolution if there is significant interdependence), but since the set of convolutions creates independent subsets for each range, it is easier to fit each range

separately. That is, we consider all delays at each range, fitting each range separately, or converting to a spectrum first if we want. The  $\chi^2$  for the fit to the entire set is equal to the sum of the  $\chi^2$ s for the subsets since the parameters for the various subsets do not interact, that is the parameters for the spectrum at one height are not required when determining the spectrum at another height. Before the deconvolution, the parameters for a particular height affect the data from a range of heights, and one would work with the whole profile at once, or at least some significant part of it.

## 8 Constrained Fitting

At the higher heights of the Arecibo World Day data, it is sometimes necessary to allow for three ions, oxygen, hydrogen, and helium. This causes no problems in the non-linear fitting process if the signal to noise ratio is good, but the highest fraction of light ions is present when the SNR is low. If all of the radar power were transmitted on a single frequency there still would be no problem, but with the power split between seven frequencies and with the spacing too close to allow the hydrogen spectrum to be resolved it becomes necessary to constrain the temperatures. This is particularly important when the fraction of  $H^+$  approaches 1 since it is impossible to determine the temperature and thus it is necessary to use information from below to derive it. The constraint technique used here is essentially the one described by Erickson and Swartz (1994). The major difference is that actual  $\sigma$ s are used in the merit function, and this makes it possible to have the constraints automatically take hold when the errors become large enough to require them. This also allows us to show rather easily that the constraints have no effect until they are large enough to smooth the temperature profile significantly; this differs from the conclusion of Erickson and Swartz (1994).

First, consider the equation for  $\chi^2$  and a modification which introduces the constraint. The model is  $y = y(x; \mathbf{a})$ , where  $x$  is the independent variable and  $\mathbf{a}$  is a vector of parameters. The  $\chi^2$  merit function is

$$\chi^2 = \sum_{j=0}^{N-1} \left[ \frac{y_j - y(x; \mathbf{a})}{\sigma_j} \right]^2.$$

Consider the following modification of the merit function:

$$\chi_c^2 = \sum_{j=0}^{N-1} \left[ \frac{y_j - y(x; \mathbf{a})}{\sigma_j} \right]^2 + \left[ \frac{a_{ip} - a_i}{\sigma_{ip}} \right]^2.$$

$a_i$  is one of the parameters, and  $a_{ip}$  is an estimate of this parameter based on other information, say a previous fit.  $\sigma_{ip}$  is a scaling parameter which measures our confidence in  $a_{ip}$ , or perhaps how much we wish to let  $a_{ip}$  influence the fit.

$\chi_c^2$  will be different from  $\chi^2$ ; this is a way of putting additional information into the fit in circumstances when the number of parameters we need to leave free, or at least partly free, is so large that we will obtain a smaller than normal average value for  $\chi^2$ , or too large a value for  $Q$ , which is derived from  $\chi^2$ . Perhaps our data are too noisy for the number of parameters we have free, but we cannot eliminate any of the parameters and still have a good model.  $\chi_c^2$  will be larger than  $\chi^2$ , and if the right level of constraint is used,  $\chi_c^2$  will have the correct value on the average. That is, the constraint will result in values of parameters different from those obtained with no constraint.  $\chi^2$  will be larger because the fit is discouraged from entering some regions of parameter space.

Now let us look at this from another viewpoint by considering the following question: how do we perform the NLLSQ fit with the modified value for  $\chi^2$ ? It turns that it can be done without modifying the fitting program, but only what goes into it. Consider  $a_{ip}$  as an additional data point associated with some value of  $x$  which is not

in use; the  $\sigma$  associated with this data point is  $\sigma_{ip}$ . We modify the model so that at this value of  $x$  it has the value  $a_i$ . Using the subscript  $c$  to describe this new data and model we have

$$\chi_c^2 = \sum_{j=0}^N \left[ \frac{y_{cj} - y_c(x; \mathbf{a})}{\sigma_j} \right]^2.$$

The derivatives associated with the new point are 1 for the parameter  $a_i$ , and zero for all other parameters. Now we fit using this augmented data set with the new function just as with the old data and function.

The claim here is that if  $\sigma_{ip}$  is set to the actual error associated with the measurement of  $a_{ip}$ , then we have included its information with the proper statistical weighting. This might seem surprising; but it is nonetheless correct. Consider how each term of  $\chi^2$  is affected by a change in  $a_i$ :

$$\frac{\partial}{\partial a_i} \left[ \frac{y_j - y(x; \mathbf{a})}{\sigma_j} \right]^2 = -\frac{2}{\sigma_j^2} [y_j - y(x; \mathbf{a})] \frac{\partial}{\partial a_i} y(x; \mathbf{a})$$

This is a statistical quantity, but we can find its expected value since  $E\{y_j - y(x; \mathbf{a})\} = \sigma_j$  assuming that we have a good fit and so the difference between the model and the data is noise. Thus

$$E \left\{ \frac{\partial}{\partial a_i} \left[ \frac{y_j - y(x; \mathbf{a})}{\sigma_j} \right]^2 \right\} = -\frac{2}{\sigma_j} \frac{\partial}{\partial a_i} y(x; \mathbf{a}).$$

This is valid for each of the  $N$  normal data points.  $\chi_c$  contains one more “data point”, and the same analysis applies, but the derivative happens to be unity in this case; thus

$$E \left\{ \frac{\partial}{\partial a_i} \left[ \frac{a_{ip} - a_i}{\sigma_{ip}} \right]^2 \right\} = -\frac{2}{\sigma_{ip}}.$$

Thus for all points, including the constraint point, changing the parameter  $a_i$  has a reasonable and consistent effect on  $\chi^2$ . The ratio of the derivative and  $\chi^2$  is independent of the scale of the magnitudes of the quantities. The constrained parameter might be very small and the other function values very large, but since the magnitude is contained in both the the derivative and the  $\sigma$ , it will not be in the ratio. The  $\sigma_{ip}$  associated with  $a_{ip}$  for a correct statistical weighting is just the error associated with its measurement. Of course, if  $a_{ip}$  is determined from some combination of previous fits,  $\sigma_{ip}$  must be computed using the correct rules of probability analysis. On the other hand, one might want to use a larger value if one wants to use a smaller degree of constraint. It might also be useful sometimes to use a smaller value to make the previous information dominate the current fit. This can be easily done since one knows the nominal value.

One can extend the above analysis to constrain several or all of the parameters.

The world day data are extremely variable in the upper ranges where the constraints might be needed, both from night to day and over the solar cycle. At the solar maximum constraints are almost never necessary, even during the winter when there is some  $H_e^+$ , because there is very little  $H^+$  and  $N_e$  is large. During the winter at night under solar minimum conditions, constraints are almost always necessary, but how much is variable.

The following procedure is suitable for these characteristics:

1. Up to some height, the fitting proceeds with no constraints.
2. We introduce the constraints beginning at the first height where three ions are allowed free. This height is not a function of conditions; the possibility of three ions is always allowed. The constraint sigmas are set so that they have very small effect a when the statistical errors are small.

3. We use the fitted values from the two previous heights to derive a linear extrapolation for the target values in the current fit.
4. As the SNR drops with increasing height, the constraint sigmas become significant and the current fit becomes dependent on previous ones from lower altitudes.
5. The degree of constraint present in any fit can be determined by looking at the temperature errors as a function of height. With no constraints present, they increase rapidly with height, but the constraints cause them to become more nearly constant. Under these conditions the temperature information is supplied from below.

When we choose the constraint sigmas, we set a target size for the errors. The fitting process then chooses the degree of dependency between the parameters at different ranges to approximately achieve these errors.

## 9 References

Erickson, P. J. and W. E. Swartz, Mid-latitude incoherent scatter observations of helium and Hydrogen ions, *GRL*, vol. 21, no. 24 pages 2745-2748, 1994

Sulzer, M. P. A phase modulation technique for a seven-fold statistical improvement in incoherent scatter data-taking, *Radio Science*, Volume 21, no. 4, pages 737-744, 1986

Swartz, W. E. Analytic partial derivatives for least squares fitting incoherent scatter data, *Radio Science*, Volume 13, no. 3 pages 581-589, 1978a

W. E. Swartz, The derivation of analytic partial derivatives for least squares fitting incoherent scatter data, Scientific Report, School of electrical Engineering, Cornell University, Ithaca NY, 14853