Ionospheric Data Assimilation Three-Dimensional (IDA3D): A global, multisensor, electron density specification algorithm

G. S. Bust, T. W. Garner, and T. L. Gaussiran II
Applied Research Laboratories, University of Texas at Austin, Austin, Texas, USA

Received 11 September 2003; revised 27 August 2004; accepted 30 September 2004; published 27 November 2004.

[1] With the advent of the Global Positioning System (GPS) measurements (from both ground-based and satellite-based receivers), the number of available ionospheric measurements has dramatically increased. Total electron content (TEC) measurements from GPS instruments augment observations from more traditional ionospheric instruments like ionospheric sounders and Langmuir probes. This volume of data creates both an opportunity and a need for the observations to be collected into coherent synoptic scale maps. This paper describes the Ionospheric Data Assimilation Three-Dimensional (IDA3D), an ionospheric objective analysis algorithm. IDA3D uses a three-dimensional variational data assimilation technique (3DV AR), similar to those used in meteorology. IDA3D incorporates available data, the associated data error covariances, a reasonable background specification, and the expected background error covariance into a coherent specification on a global grid. It is capable of incorporating most electron density related measurements including GPS-TEC measurements, low-Earth-orbiting "beacon" TEC, and electron density measurements from radars and satellites. At present, the background specification is based upon empirical ionospheric models, but IDA3D is capable of using any global ionospheric specification as a background. In its basic form, IDA3D produces a spatial analysis of the electron density distribution at a specified time. A time series of these specifications can be created using past specifications to determine the background for the current analysis. IDA3D specifications are able to reproduce dynamic features of electron density, including the movement of the auroral boundary and the strength of the trough region.

INDEX TERMS: 2447 Ionosphere: Modeling and forecasting; 2494 Ionosphere: Instruments and techniques; 6982 Radio Science: Tomography and imaging; 6969 Radio Science: Remote sensing; KEYWORDS: data assimilation, tomography, Kalman filter, objective analysis


1. Introduction

[2] With the advent of slant total electron content (TEC) measurements from dual-frequency Global Positioning Satellite (GPS) receivers, the volume of ionospheric data available for scientific analysis has greatly increased. When these measurements are combined with more traditional ionospheric observations (e.g., ionospheric soundings, in situ satellite measurements), several thousand, even tens of thousands, of observations of the ionospheric electron density distribution can be collected within a given 15 min period. These data represent both point measurements of the density and integrated measurements of electron content. They are distributed around the globe with a concentration in North America, Europe, Japan, and Australia. While some of the measurements are coincidental, many are complementary. For example, TEC measurements from a ground-based GPS receiver contain information on the ionosphere as well as the plasmasphere and at times the TEC measurements can be dominated by the topside ionosphere and plasmasphere. Ionosondes measure the electron density below the $F$ region peak. By combining these data sets in a consistent manner, the electron density profile can be specified to higher altitudes, providing a better topside specification. In a similar manner, a collection of geographically dispersed measurements can be combined into a single picture of the large-scale ionospheric behavior [e.g., Garcia-Fernández et al., 2003a, 2003b].

[3] While this infusion of large amounts of data creates new and exciting opportunities for ionospheric physics, it also raises several issues: (1) What is the optimum way to combine these different data sets, each with its own sources of error, into a consistent synoptic or global specification? (2) How important are the various data sources to the overall global specification of electron density? (3) How
accurate is the resulting electron density specification and what is its error? [4] As the ionospheric community confronts these issues, it can rely on the techniques developed by the meteorological and oceanographic communities. Since the meteorological community developed the initial techniques, they also developed the language. This paper will use the meteorological terminology and attempt to explain these terms in the ionospheric context. A more complete description of the terminology and techniques used by meteorologists can be found in the works of Daley [1991], Courtier et al. [1997], Lorenc [1986], Tarantola [1987], and Menke [1989]. The term “analysis” will be used for a large-scale specification based upon a collection of different data. In meteorology, an analysis is also called a weather map or chart. By analogy, an ionospheric analysis is a space weather map. The term “objective analysis” is used to describe an analysis generated through an automated process. An analysis generated through the active participation of a human scientist is called a “subjective analysis.” A “spatial analysis” is an analysis that completely specifies the spatial weather at a given time. It is a snapshot of the weather. The term “statistical minimization” is used to describe any numerical technique that minimizes a cost function of the statistical values. The meteorological objective analysis algorithms were first developed in the late 1950s [e.g., Panofsky, 1949] and are now well developed. In the past 50 years the meteorological community has developed numerous mathematical techniques for performing objective analysis (see Daley [1991] and Menke [1989] for a survey of these techniques).

[5] An ionospheric objective analysis algorithm has several requirements that it must satisfy. It must be able to ingest measurements from different types of sensors and to use any measurement that can be derived from the electron density, the instrumentation, and the observation geometry. Because of the different types of data and their geometry, the algorithm must be three-dimensional. It also must weight the influence of the data sources. In addition, the objective analysis algorithm must have a mathematically rigorous way of determining the extent of data influence into regions where there is no data and must smoothly fall back to a predictive model far from data. Finally, it should be usable as an input to a data assimilative model.

[6] The development of an ionospheric objective analysis algorithm is important in developing a full (driven by a full-physics model) data-assimilative (time-dependent), ionospheric forecast. The standard data assimilation cycle [Daley, 1991] is a four step cycle that includes quality control, an objective analysis (to generate a complete spatial field with the available data), model initialization, and a theoretical forecast (to propagate the spatial field forward in time). The necessary components in this cycle are quality control algorithms, an objective analysis algorithm, and a forward predictive model. Model initialization is the essential process of data assimilation, while the other components can be viewed as input algorithms. A good data assimilation model should develop an initialization scheme that allows for new and different quality control, objective analysis, and/or predictive models to replace existing algorithms seamlessly.

[7] Instead of ingesting data directly into the predictive model (i.e., replacing the model output with an observation at only the points of the observation), observations are projected by an objective analysis algorithm into the proper scale and onto the proper grid for the predictive model and combined with the existing model prediction. The weather map created by the analysis algorithm replaces the model output at the previous time step during the initialization phase, and the physics model moves forward in time. The data assimilation cycle exist for operational, numerical, and physical reasons. Operationally, nowcasts and forecasts will not be useful if the predictive model needs to wait for all of the available observations to be collected and quality checked. Numerically, the computational resources to both assimilate each of the individual measurements and to calculate the future state of the system are immense. As both a physical and numerical reason, introducing individual datum into the predictive model can generate artificial discontinuities that propagate through the model system. These discontinuities can be avoided by collecting the data into a spatially continuous structure that is ingested into the predictive model at the same time. While this procedure can be performed within a predictive model, it is still a separate algorithm that can be treated independently of the physics model. Because of its role with the data assimilation cycle, objective analysis algorithms need to reproduce the background model predictions far from data and produce corrections in data-rich regions that are as free of spurious nonphysical modes as possible.

[8] Several different techniques have been developed to generate objective analysis of the ionospheric electron density. One of the earliest techniques is computerized ionospheric tomography (CIT). CIT is a remote sensing inversion technique that in standard usage, develops a two-dimensional electron density specification from a series of one-dimensional ionospheric observations (typically line-of-sight column electron density) and various minimization criteria [Austen et al., 1988; Kersley et al., 1993; Raymund et al., 1994; Bust et al., 1994; Kronschnabl et al., 1995; Raymund, 1995]. As such, CIT is a spatial analysis technique. Standard CIT techniques use a single array of receivers, normally aligned in latitude along a constant longitude, and a low-Earth-orbiting (LEO) “beacon” satellite as the transmitter source (typically in polar orbit). For LEO satellite orbits, whose orbit longitude almost coincides with the array longitude, the collected data can approximately be considered to be in a two-dimensional plane defined by the latitude extent of the array and altitude to the satellite. For such an alignment the standard mathematical techniques for limited angle tomography apply, and CIT produces two-dimensional electron density maps in the region of the receiver array. This leads to three limitations for standard LEO-based tomographic methods. First, the reconstructed electron density is only two-dimensional. For satellite orbits offset by a large amount from the receiver array, the fidelity of the tomographic inversions degrades significantly. Second, the ionosphere is considered static over the time period of the data collection (typically ~20 min). Third, the satellite passes do not occur continuously in time. For example, a midlatitude receiver array will collect data from ~15 to 20 satellite passes per day with the passes spaced irregularly in time over the day. Thus
traditional CIT cannot suitably produce a global spatial analysis of the electron density at regular update times. If there were enough satellite sources and ground receivers distributed globally to make CIT practical, CIT would be a useful global objective analysis method. Currently, even with the addition of a global network of ground GPS stations, the data is too sparse for CIT techniques to be applied directly. Thus different numerical methods are needed to produce the desired specification.

To overcome the limitations of CIT and to exploit other data sources (especially GPS-TEC measurements), other numerical techniques have been developed. Regional specications of electron density using combined data sets of CIT, ground GPS, and ionosondes have been developed at ARL:UT to overcome these and other weaknesses [Coker, 1997; Kronschnabl et al., 1997]. When compared with independent sources, the results from these regional specication algorithms have been promising. However, the methodologies employed are somewhat ad hoc and do not allow for a mathematically rigorous method of adding arbitrary data sources. Similar work by other groups have led to the development of more advanced spatial analysis techniques. Fremouw et al. [1992] developed a direct inverse theory [Menke, 1989] approach that made use of empirical orthogonal functions (EOF). Howe et al. [1998] built upon the work by Fremouw and developed a Kalman filter method for the ionosphere based upon spherical harmonics horizontally and EOFs vertically. Hernández-Pajares et al. [1999, 2002] and García-Fernández et al. [2003a, 2003b] have developed a voxel-based Kalman filter of the vertical electron content using GPS-TEC measurements from ground receivers and occultation satellite receivers and ionosondes. Recently, Schunk et al. [2004] have described a Global Assimilation of Ionospheric Measurements (GAIM) cycles. They have a Gauss-Markov Kalman filter currently implemented and a full physics-based Kalman filter under development. Wang et al. [2004] describe a Global Assimilative Ionospheric Model (GAIM). They have a physics-based Kalman filter implemented and a 4DVAR method under development.

This paper describes the development of the Ionospheric Data Assimilation Three-Dimensional (IDA3D) algorithm, an ionospheric objective analysis algorithm. IDA3D creates a global three-dimensional electron density specification by ingesting ionospheric measurements from a variety of instruments. IDA3D builds upon and extends previous work at ARL:UT in computerized ionospheric tomography (CIT) and in combining CIT, GPS-TEC, and ionosonde data to produce regional specications of electron density [Bust et al., 1994; Kronschnabl et al., 1995; Coker, 1997; Kronschnabl et al., 1997]. While in many ways IDA3D is an extension of this previous work, it treats and handles both observational data and errors in different ways and it is based upon a fundamentally different mathematical technique than the previous CIT algorithms developed at ARL:UT.

The paper is organized as follows. Section 2 introduces three-dimensional variational data assimilation (3DVAR), the numerical method upon which IDA3D is developed. Section 3 contains a mathematical description of the IDA3D algorithm. Section 4 describes the data sources currently used by IDA3D. Section 5 presents sample results from IDA3D to demonstrate how it works in the field upon actual data. Finally, section 6 provides a discussion of some of the outstanding issues regarding objective analysis for the ionosphere and IDA3D as well as some future plans for improvements of the algorithm.

2. Three-Dimensional Variational Data Assimilation (3DVAR)

Traditional CIT methods can be considered a subset of statistical minimization methods. These minimization methods are capable of taking any ionospheric measurement linearly related to electron density (or nonlinear data such as EUV limb-scan observations) and optimally estimates an electron density field from the measurements. The method 3DVAR [Daley, 1991; Courtier et al., 1997] is a statistical minimization method that seeks to minimize a cost function of data perturbations weighted by the data error covariance and the deviations of the model from the background weighted by the a priori background model error covariance. The background error covariance is obtained from the background error \( \tilde{\beta} = \tilde{x}_b - \bar{x} \), where \( \bar{x} \) is the “true” electron density and \( \tilde{x}_b \) is the electron density from the background model. Assuming the background error is unbiased, the background error covariance matrix can be written as \( P_b = \langle \epsilon_b \epsilon_b^T \rangle \). In a similar fashion, the data vector, the data error, and the corresponding data error covariance matrix can be defined as \( \tilde{y}, \tilde{d}, R = \langle \epsilon_d \epsilon_d^T \rangle \), respectively. The data error is assumed to be unbiased, and usually the data errors are assumed to be independent, making the data covariance diagonal.

The 3DVAR method naturally lends itself to multiple types of data, provides a self-consistent method of estimating the state variables, allows all the observations to influence the analysis at every grid point, and allows more general forward models to be used. Thus observations that are not directly related to the electron density (such as TEC) can be more easily assimilated [Daley and Barker, 2000]. One of the important characteristics of this algorithm is that all the a priori knowledge required to obtain a unique specification is carried in the background model predictions and the error covariances (both the model and data). All three of these quantities are physical quantities that are, at least in principle, derivable from the basic physics of the ionosphere. For a full mathematical treatment of 3DVAR, see Daley and Barker [2000], Daley [1991], Tarantola [1987], Lorenc [1986], Heckley et al. [1992], Parish and Derber [1992], Courtier et al. [1997], and Cohn et al. [1998].

The 3DVAR technique offers several features that make it useful for ionospheric specification. First, such a method naturally lends itself to specifying the electron density field on a three-dimensional grid, where the grid can be entirely irregular if desired. In 3DVAR, the relationship between the influence of data and initial model on the 3-D grid is determined by the error covariances, particularly the spatial correlation lengths.

A second feature of the method is the simplicity of adding additional data sources. All that is required is a mathematical relationship between the measurement and electron density and a specification of the data covariance. The method automatically weights the data types against
other data types and the initial background model when estimating the electron density.

[16] A third feature of this algorithm is that it is interpolative. That is, it interpolates (in a mathematically rigorous way) between model prediction and the influence of the data. In data-rich regions the solution is determined by the data. If several data sources are in the same region, the relative weights (given by the error covariance) are used to determine the influence of the data. Far from data, beyond a correlation length, the result smoothly returns to the background model estimate. Thus there is a well-defined transition from data-rich regions to data-poor regions, and the transition is governed by the relative weights of the data and background model error covariance. In addition, the covariance weighted transition makes it straightforward to determine which regions of the estimated electron density are driven by data and which are driven by the background model. This property is important for systems that must always have a reliable estimate of electron density (and an estimate of the error).

[17] This leads to the fourth feature. The objective analysis technique produces a formal error estimate as part of the minimization method. The formal error (formal since, in general, the error covariances are not known exactly), when combined with the knowledge of what regions were strongly driven by data, is useful for applications that need to be able to estimate the accuracy of their application performance.

[18] Finally, the fifth feature is the basic physical principles underlying the algorithm. All the a priori information of the objective analysis method is carried in the data error covariance, the background model, and the model error covariance. The covariances are, in principle, determinable from either physical modeling or experimental measurements. Such grounding in basic physical principles is important for any complex remote-sensing inversion algorithm and is not the case for the various methods that have been used to develop traditional CIT algorithms such as maximum entropy [Fourgerre, 1992] and Tiakov regularizaton [Fehmers, 1996].

3. Description of the IDA3D Algorithm

3.1. Derivation of IDA3D Algorithm

[19] To derive the algorithm, we begin at some time \( t \), with a background array of model (either theoretical or empirical) electron density values \( \tilde{x}_b \) on a model grid. We wish to obtain an array of analyzed electron density values \( \tilde{x}_a \) on the same grid by using measured data. The modeled electron density is represented as voxel values on a three-dimensional grid. The density is constant within each voxel. The data values \( \tilde{y} \) are related to the model electron density through a forward operator \( \tilde{H} \). For direct measurements of density such as an ionosonde or in situ satellite measurements, the forward operator is an interpolative operator that interpolates the model density from the grid location to the measurement location. For TEC, the equation relating the data to the electron density is given as

\[
T(\tilde{r}_s, \tilde{r}_r) = \int ds N_e(\tilde{r}),
\]

where \( T \) is the electron content from the satellite position \( \tilde{r}_s \) to the receiver position \( \tilde{r}_r \) and \( N_e \) is the electron density. Representing the electron density as a field of values on a three-dimensional grid and taking a discrete approximation of (1) produces

\[
\tilde{T} = \tilde{H}\tilde{N}_e. \tag{2}
\]

Then, the forward operator becomes a matrix \( \tilde{H} \) and is generally referred to within the ionospheric tomography community as the geometry matrix. Here \( \tilde{H}_{ik} \) relates the electron density in voxel \( k \) to the TEC of data \( i \). In the simplest form, that relation is the length of the ray in the voxel. Because of uncertainties in the receiver bias for GPS data, TEC measurements are often treated as relative, rather than absolute measurements. In this treatment a reference measurement for each receiver is subtracted from the TEC measurements to remove the receiver bias. This treatment of the relative TEC introduces off-diagonal elements in the covariance matrix, which are not presently treated. This changes equation (1) to

\[
T_{rel}(\tilde{r}_s, \tilde{r}_r) = \int ds N_e(\tilde{r}) - T_0(\tilde{r}_s, \tilde{r}_r)
\]

and (2) to

\[
\tilde{T}_{rel} = (\tilde{H} - \tilde{H}_0)\tilde{N}_e, \tag{4}
\]

where \( \tilde{T}_{rel} \) is the relative TEC measurement, \( T_0 \) is the measurement of the designated reference satellite for the given receiver (typically the smallest TEC measurement), and \( \tilde{H}_0 \) is forward operator representing the ray path of the reference measurement. This relationship holds for any measurement that is related to electron density through a line integral. For simplicity in the derivation, in the following it is assumed that the forward model \( \tilde{H} \) consists of both the direct forward operator and the reference operator as described in equation (4).

[20] Using the definitions of \( \tilde{H} \), the model error covariance \( \tilde{P}_b \), and the data error covariance \( \tilde{R} \) described in section 2 and assuming that the background and data errors are distributed normally, then the maximum likelihood estimate for the analyzed electron density field \( \tilde{x}_a \) is obtained by minimizing the scalar cost function \( J \) with respect to \( \tilde{x}_a \), where

\[
J = \frac{1}{2} [\tilde{y} - \tilde{H}\tilde{x}_a]^T \tilde{R}^{-1} [\tilde{y} - \tilde{H}\tilde{x}_a] + \frac{1}{2} [\tilde{x}_b - \tilde{x}_a]^T \tilde{P}_b^{-1} [\tilde{x}_b - \tilde{x}_a]. \tag{5}
\]

The first term in equation (5) represents the influence of the data. If this was the only term, equation (5) would represent a classic \( \chi^2 \) minimization problem. The data error covariance includes instrumental errors and errors of representation. These representation errors are caused by observations of the subscale phenomena and the discretization of the spatial grid. These errors have to be included to provide an accurate data error covariance. For example, if the analysis period is 15 min and all GPS data from a single receiver-satellite is averaged over that period, the fluctuations of the data about that average is an error of representation that must be added to the data error.
[21] The second term represents the influence of the model (or background) field, \( \tilde{x}_b \), with the weight of the influence given by the model covariance \( \tilde{P}_b \). It is interesting to note that the second term has the same functional form as techniques that invoke Tikhonov regularization [Fehmers, 1996], with the inverse of the model covariance taking the place of the smoothing matrix. Thus, in some sense, this form of the cost function can be considered a more sophisticated form of "regularization."

[22] The solution to equation (5) can be derived in "model" space where the relevant matrix scale is given by the number of unknowns or in "data" space where the number of data determines matrix size. Since the number of unknowns (i.e., the electron density at every grid point on a 3-D grid) will typically far outnumber the number of measurements, IDA3D solves equation (5) in data space. The solution, in data space, [Daley and Barker, 2000] for the analyzed electron density \( \tilde{x}_a \), in equation (5) is then given by

\[
\tilde{x}_a - \tilde{x}_b = \tilde{P}_a \mathbf{H}^T [\mathbf{R} + \tilde{H} \tilde{P}_b \tilde{H}^T]^{-1} (\tilde{y} - \tilde{H} \tilde{x}_b),
\]

where \( \tilde{y} \) is the vector of observations related to the electron density. In addition to the analysis solution, the formal error covariance on the analysis is given by

\[
\tilde{P}_a = \tilde{P}_b - \tilde{P}_a \mathbf{H}^T [\mathbf{R} + \tilde{H} \tilde{P}_b \tilde{H}^T]^{-1} \tilde{H} \tilde{P}_b,
\]

with \( \tilde{P}_b \) referring to the background error covariance and \( \tilde{P}_a \) the analysis error covariance. The diagonal elements of equation (7) are the analyzed variances. The algorithm currently only calculates the analyzed variances since the full calculation is computationally prohibitive. When sufficient computational resources become available, IDA3D will calculate an approximation of the full matrix.

[23] Equations (6) and (7) have been implemented using the GPS ground-based TEC, GPS occultation TEC data, and CIT-beacon TEC sources, as well as ionosondes and in situ satellite data sources. Currently, IDA3D has used the International Reference Ionosphere (IRI) [Bilitza et al., 1993; Bilitza, 2000], the Parameterized Ionosphere Model (PIM) [Daniell et al., 1995], the RIBG [Reilly, 1993; Reilly and Singh, 1997] and the global core plasma model [Gallagher et al., 2000] empirical models as background specifications. We have recently implemented the Advanced Space Environment (ASPEEN) version of TIMEGCM [Roble and Ridly, 1994], a fully coupled ionosphere-thermosphere predictive physics model as the background model to IDA3D.

3.2. Background Model and Error Covariance

[24] Any empirical or predictive ionospheric model can be used as the background or initial guess model for IDA3D. The current version of the algorithm allows the user to input the background model that is typically an empirical model. Other models will be used in the future.

[25] To model the background model error covariance, we make the following assumptions. (1) Spatial correlations are separable horizontally and vertically. (2) Vertical correlations are given by a Gaussian. (3) Horizontal correlations are given by an elliptical Gaussian in geomagnetic coordinates.

[26] With the above assumptions, the error covariance between grid points \( k \) and \( l \) is given as

\[
\langle \tilde{P}_{kl} \rangle_{\mathbf{w}} = (S_b)^{1/2} (S_b)^{1/2} e^{-\gamma_{kl}/L(\alpha \beta 2)},
\]

where \( S_b \) is the background model variance, \( z \) is the height, \( \gamma_{kl} \) refers to the great circle distance between points \( k \) and \( l \), and \( L_z \) is the vertical correlation length. The azimuthally dependent, horizontal correlation length \( L(\alpha) \) is given as

\[
\frac{1}{L(\alpha)^2} = \frac{\sin^2(\alpha)}{L_\phi^2} + \frac{\cos^2(\alpha)}{L_\theta^2}.
\]

For both latitude (\( \theta \)) and longitude (\( \phi \)), the correlation length is a function of the geomagnetic latitude, magnetic activity, and time of day. The geomagnetic latitude dependence is obtained by providing separate correlation lengths specified by the user for low latitudes, midlatitudes, and high latitudes (described below). A smooth transition from one region to the other is obtained by using a transition function

\[
T(\theta_0, \phi_0, d\theta) = \frac{1}{1 + \exp(-(\theta - \theta_0)/d\theta)},
\]

where \( \theta_0 \) is the starting latitude of the transition region and \( d\theta \) is the width of the transition region. Given a set of three parameters \( (L_i, L_m, L_h) \) defining the low, middle, and high latitude correlation lengths, the latitude model of the correlation length is given as

\[
L_\theta(\theta) = L_i + (L_m - L_i) T(\theta, \theta_0, d\theta) + (L_h - L_m) T(\theta, \theta_0, d\theta),
\]

and a similar relation holds for the geomagnetic longitude \( L_\phi(\phi) \). Finally, for two points \( (\theta_1, \phi_1) \) the \( L_{\phi\theta} \) term in equation (9) is given as \( L_{\phi\theta} = L_{\phi\theta}^i L_{\theta\theta}^i \), with a similar relation holding for \( L_{\theta\phi}^i \). The low-latitude and high-latitude transition positions are dependent on the magnetic activity level and are given through empirical functions of \( Kp \). The latitude and longitude correlation parameters \( (L_i, L_m, L_h) \) are defined in the user configuration file.

[27] The vertical correlation lengths are given as either a vector of lengths in kilometers (typically plasma scale heights) or as a vector of number of nearest neighbors in grid spacing and are defined in a user configuration file. The vertical correlation length is determined by the value of the correlation scale at each height \( L_z(z_i, z_i) = (L_z(z_i) L(z_i))^{1/2}, \) with \( z_i \) and \( z_i \) representing the two heights. While currently the correlation lengths are given through the user configuration file, future enhancements will allow the scale heights to be derived from plasma temperature either obtained from a model or observations.

[28] The horizontal and vertical error correlation lengths of the ionospheric weather are not well known (the climate values are better understood). This is particularly true for geomagnetically active times and at high or low latitudes. A few studies have been conducted to estimate the midlatitude
quiet time horizontal correlation lengths [Rush, 1976; Gail et al., 1993; Bust et al., 2001], and these values provide the default correlation lengths used in IDA3D. Since all the parameters are configurable by the user, as better estimates for the correlation lengths and boundary regions become available, they can be incorporated into IDA3D runs. Typical correlation lengths for latitudes are 3 degrees at low and high latitudes and 5 degrees at midlatitudes. For longitudes the lengths are given in great circle degrees, and typical values are 5 degrees at low latitudes, 10 degrees at midlatitudes, and 4 degrees at high latitudes. In altitude, the correlation length varies from 20–25 km in the E and F regions to 500 km in the plasmasphere. Thus how realistic the correlation model parameters are (both vertically and horizontally) is only limited by how well the correlation lengths are known as a function of geomagnetic region, time of day, solar cycle, and magnetic activity.

3.3. Numerical Grid

[29] IDA3D can be run on a region of the ionosphere or on a global grid. The data available to IDA3D is distributed nonuniformly over the globe. Ground GPS data has a higher density of receivers in the USA (particularly California), parts of Europe, and Japan than at other regions of the globe and has very little coverage over the oceans. Regional data sets such as those obtained from LEO “beacon” tomography data provides very high resolution information over a small spatial region. Thus to accurately characterize the distribution of data, as well as for computational reasons, we use an irregular grid in the horizontal dimension. Currently, the user can specify that the horizontal grid be computed within the algorithm or from a precomputed grid file. The precomputed grid has a higher density of grid points near regions of large data coverage, and we have a regionally dense grid where beacon tomography arrays are located (currently Alaska and Greenland). Owing to computational limitations, the standard precomputed grid used for the single workstation version of IDA3D has a default spacing of 3° of latitude. The default longitude spacing and number of grid points depends on the latitude. At the equator the spacing is 5°, giving 72 points. At higher latitudes we maintain a spacing of 5° great circle distance. Thus for a given latitude, the number of longitude points is given by \( \sim 72 \cos(\theta) \), where \( \theta \) is the latitude. At higher latitudes there are fewer longitude grid points. The fact that the number of longitude points is a function of latitude is what makes the horizontal grid irregular. For the default spacing described above, we get \( \sim 3000 \) total horizontal grid points. In addition, in regions where there are tomography arrays, the horizontal grid can be embedded with a high-resolution regional grid that can increase the total number of horizontal grid points significantly. Typically, the vertical grid consists of 40–100 elements from 90 km to several thousand kilometers (depending on the application). Vertical grid spacing is typically \( \sim 10 \) kilometers in the E region, \( \sim 20 \) km in the F region, and increasing on the topside and into the plasmasphere to several hundred kilometers. Depending on the various resolutions chosen by the user and what altitudes the grid extends, the total number of grid points can vary from \( \sim 100,000 \) to several hundred thousand points. Single workstation runs of IDA3D grids larger than \( \sim 100,000 \) points are computationally very intense. However, the multiprocessor version of IDA3D currently being developed will not have the computational limitations of the workstation version, and it will typically use a much higher resolution grid with higher resolutions in regions with high data density.

3.4. Temporal Updates to the Background Model

[30] Our motivation is to produce the most accurate spatial analysis possible at the analysis time and improve our initial background model of the electron density as time proceeds. There are currently three methods which are used in IDA3D. These methods are chosen in the run-time configuration file. One is to not update at all, to use the climate, and to allow IDA3D to correct it. This has a certain appeal as it directly allows one to investigate where the data is adding “weather” to the climate. Unfortunately, the background climate model is often very far from the weather, and IDA3D takes much longer to converge to a good solution. In order to increase computationally efficiency, IDA3D often uses previous weather maps as a background model. A second method is to insert the previous analysis as the background field and not update the covariance. This has the advantage of simplicity and, for short temporal updates, will work quite well. The third option is to use a Kalman filter for the temporal updates. This requires a propagation model. Since we are not doing a full predictive first principle model, we apply a first-order Gauss-Markov filter [Gelb, 1974; Howe et al., 1998]. A Gauss-Markov process is not the only propagation model possible. A linear prediction [Press et al., 1992] model has been applied with success to prediction of ocean wave heights and is something we may pursue in a future version of IDA3D.

4. Current IDA3D Data Sources

[31] As described in section 3, IDA3D ingests several types of ground-based and space-based measurements. The ground-based CIT TEC data are high-resolution regional data sets, while the ground GPS TEC data, satellite occultation GPS TEC data, satellite in situ electron density measurements, and ionosonde measurements are global lower resolution measurements.

[32] The ground-based TEC data are given as slant TEC measurements (1) and are taken from the global network of GPS receivers and from arrays of CIT beacon receivers. The satellite occultation data of TEC is similar to the ground TEC measurements but with differing look angles. Measurements of electron density are taken from the global network of ionosondes and in situ satellite measurements.

[33] The ground GPS TEC data and ionosonde data are available globally and at regular intervals in time. The time interval for GPS data is typically 30 s, while the time interval for ionosonde data is typically 15–60 min. The beacon array TEC data are regional in nature and are available only irregularly in time. The sampling time for the beacon TEC is determined by satellite geometry, number of satellites, and placement of receivers. The coherent ionospheric Doppler receivers (CIDR) developed...
at ARL:UT are able to track up to three satellites simultaneous-ly, and there are currently nine satellites available. At low latitudes the CIDR receivers will see approximately 18–20 passes per day, while at high latitudes they can see 50 or more passes per day. These different data types are complementary. While beacon TEC data produces very high resolution images of electron density that are regional, they are limited by the array geometry and occur irregularly in time. The GPS TEC data produces lower spatial resolution electron density information that is global and available regularly in time.

[35] The satellite data are complementary to the ground instruments. GPS occultation satellites produce several hundred occultations per day. Over a several day period, the data is distributed over the entire globe. Thus over long timescales the occultation data is available on a global basis. However, at any given time, the data is localized in space and acts more as a high-resolution regional data set. Similar arguments hold for satellite in situ measurements of electron density.

[35] Raw data must be preprocessed and quality controlled before it is ready for the IDA3D algorithm. Some of the data sources will be collected at a higher spatial and/or temporal resolution than the IDA3D grid. For self-consistency, these data are averaged or subsampled to a resolution consistent with the IDA3D grid. While in principle, the algorithm should take care of the averaging in a least-squares sense (through the correct use of representation error), there are several reasons (some physical and some computational) for averaging down the data. First, data points too close together may have errors that are linearly dependent if the ionospheric variation is below the noise level (such as if often the case with the southern California GPS receiver network). This linear dependence increases the condition number in the matrix to be inverted, and can produce numerical instabilities. Second, sometimes the subgrid variations are huge, such as in situ electron density measurements within an equatorial bubble. This leads to problems with our quality control algorithms that use the standard deviation as a rejection scheme. Finally, averaging reduces the size of the data array and hence the computational time.

[36] The raw data that produces beacon TEC and GPS TEC (both ground and occultations) are subject to phase cycle slips. These cycle slips must be detected, corrected for, and flagged as bad data prior to ingestion into IDA3D. The measured TEC is only known up to a constant that depends on the satellite and receiver. For beacon TEC and GPS occultation data, one data point from each receiver is used as a reference and subtracted from the other data to eliminate the unknown constant. For GPS TEC, existing independently estimated satellite bias estimates are used, and one of the satellite TEC measurement from each receiver is subtracted off as a reference to eliminate the receiver bias. Ionosonde data has measurement errors in the density estimates that can be due to bad ionograms (lots of noise) or the scaling method (if autoscaling is used). The data is required to fall within certain minimums and maximums, and some manual plotting is done for general sanity checks. DMSP electron density measurements have their own data quality flags which are used to filter the data. In addition, the data must fit minimum and maximum bounds in a similar manner to the ionosondes. A suite of preprocessing and quality control algorithms have been developed for each data type ingested by IDA3D.

4.1. Beacon CIT Data

[37] Beacon CIT arrays calculate relative total electron content (TEC) data from the observed Doppler shift at 150 and 400 MHz emitted by low-Earth-orbiting (LEO) satellites. Each receiver can be considered as a separate instrument. Each array can be considered as an extended instrument. For the latter case the observation period across the array lasts for ~15–20 min, which becomes the measurement time of the array.

[38] When the CIT array data is ingested into IDA3D as if a single extended instrument, it provides high-resolution two-dimensional information horizontally (along the array direction) and vertically to the analysis algorithm. The horizontal resolution obtainable with CIT arrays can be as small as a few kilometers, while experimental investigations [Watermann et al., 2002; Coker et al., 2001] have demonstrated that the vertical resolution is ~20–50 km (limited by the minimum elevation angle). When used in this fashion, CIT represents an inexpensive unique data source that provides high resolution spatial imaging of the ionosphere. The electron density analyzed from CIT arrays is similar to the types of images obtained from incoherent scatter radars when they are in elevation scan mode.

[39] There are currently CIT arrays located in Alaska, Greenland, Scandinavia (near the EISCAT ISR), and in Massachusetts near the Millstone Hill ISR. There is a single receiver in Ancon, Peru, and another currently being tested in Bogata, Columbia. Plans are underway to put four additional receivers in South America in an East-West configuration in support of the C/NOFS satellite mission. There are seven beacon satellites suitable for CIT in orbit. The Navy Ionospheric Monitoring System (NIMS) has four satellites in orbit, and there are beacon transmitters on PICOSat, Radcal, and GFO. In the near future, the COSMIC constellation of six satellites will launch, each with a beacon transmitter. Finally, C/NOFS will have beacon transmitters available.

4.2. GPS Data

[40] There are more than 1000 dual-frequency GPS ground stations available globally. Our standard automated download and processing looks at three FTP sites for GPS data (NASA, available at cddisa.gsfc.nasa.gov, UCSD, available at lox.ucsd.edu, and NOAA, available at www.ngs.noaa.gov). Each station tracks between 8 and 12 satellites at a time and collects data at a rate of once every 30 s or faster. GPS receivers with dual frequency phase data are suitable for TEC measurements and can be used as source data for IDA3D. GPS provides data continuously on a global basis, but the satellites move across the sky slowly, with ~100 m/s speed at the F region intercept of the ionosphere. The slow movement of satellites, combined with the small number of satellites in view at a given time limits the vertical resolution of ground GPS. Typically, ~500 GPS stations provide good data (generally a large number of stations do not survive the processing and quality control due to bad data for that station for that day. In addition, stations that are within the same grid cell and are
rejected for redundancy). They each track ~8 satellites, making ~4000 measurements distributed over the globe at a given time. If this data was uniformly distributed, each data point would cover approximately a 3° by 3° horizontal patch of the ionosphere. However, each measurement is a slant TEC measurement that would still distorting the distribution and blurring the vertical distribution. In addition, the ground GPS data is not uniformly distributed over the globe. Thus in regions with a large number of ground stations, GPS can provide good horizontal resolution with fair vertical resolution. While the data is integrated electron density along the receiver-satellite path (much the same as the beacon CIT data), it is unsuitable for a direct tomography inversion due to the spatial sparseness of the data. Hence the ground GPS data provides good horizontal information and weak vertical information on electron density, continuously in time over the entire globe and is complementary to the beacon CIT data.

4.3. Ionosonde Data

There are more than 40 ionosonde stations that regularly report to the Space Environment Center (SEC). Many of the stations provide a full set of ionosonde scaling parameters including $E$ region, $F_1$ region, and $F_2$ region measurements of peak electron densities, peak heights, and layer thickness. The data is globally distributed and, up to some sample interval time, run continuously. The ionosonde data provides a good complement to the ground GPS TEC data. These data provide direct point measurements of layer peak values electron density (primarily $E$ layer and $F_2$ layer) over the globe. However, the heights, particularly the height of the $F_2$ peak, are not direct measurements but are taken from various inversion schemes and are subject to increased error. This is also true for retrieved profiles. Quite often the inversion techniques work very well and the retrieved heights are quite accurate. However, it is important to keep in mind from a data quality point of view the heights are not measurements with instrumental error. They are inversion products with an error dependent on the data quality and inversion process. Presently, only the $E$ and $F_2$ peak densities and estimated heights are used in IDA3D, not the entire profile. The errors in the densities are increased to account for errors in the estimation of the peak height.

4.4. Satellite In Situ Data

Satellites can provide in situ measurements of electron density along the satellite trajectory. Currently, the in situ electron density provided by DMSP is easily available and is ingested into IDA3D. The data is global in the sense that over time all parts of the globe is sampled. However, over an IDA3D assimilation time period, the data only samples a small spatial region. For a given IDA3D analysis, the data provides regional high-resolution direct measurements of electron density. By providing measurements of electron density at a constant altitude on the topside, satellite in situ data complements the TEC data and provides part of the topside profile to the IDA3D solution.

4.5. GPS Occultation Data

Currently, IDA3D receives GPS TEC occultation from satellites carrying GPS occultation receivers including the CHAMP, SAC-C, PICOSat, and GRACE satellites. While ground slant TEC data provides excellent horizontal information on ionospheric electron density, GPS occultation data provides very good vertical information on electron density. IDA3D accepts the relative TEC from the occultation satellite. That is, the TEC is known up to an initialization constant. The TEC provides an integral over the entire ionospheric path between the receiver and GPS satellites. While a large part of the TEC information is given by the height at the tangent point, a significant amount of information is given from all heights along the path. This is particularly true for paths with the tangent height in the $E$ region. The $F$ region is traversed twice and provides a significant fraction of the total TEC. Since IDA3D uses a global 3-D grid, it is able to sort out the various contributions along the path. It can provide a more accurate representation of electron density than an inversion method that assumes a spherical ionosphere [Hajj et al., 2000]. The coverage of the occultation data is similar to that provided by in situ spacecraft. The data coverage is global, given enough time for the satellite to cover the globe. However, for any given small sampling period the occultations are localized in space.

5. Sample IDA3D Results

As an example of how multiple data sets can be ingested into IDA3D on a global basis, IDA3D was run for 12 December 2001. The algorithm was run in a historical mode (significantly after the date) with a 15 min update time. Approximately 300 GPS ground receivers were available for the run, as well as over 40 ionosondes. GPS occultation data was available for the entire day from the CHAMP satellite and available from 0700 to 1200 UT from PICOSat. In addition to the above globally available data sets, data was available from two high-resolution tomography arrays. One array, located in central Alaska, consisted of five receivers extending from 60° to 70° of geographic latitude. The second array, consisting of five receivers, is located along the west coast of Greenland extending from ~60° to 77° of geographic latitude.

To get a sense of how the data is distributed globally, Figure 1 shows a representative coverage plot for 0830 UT. Notice the very high data coverage of GPS in North America and Europe. However, the GPS coverage is relatively sparse at high latitudes owing to the orbital inclination of the GPS constellation. One advantage of CIT data is the polar orbits of the satellites, thus providing additional high-altitude data. Although there are over 40 ionosondes available, not all of them have good quality data all the time. In addition, some ionosondes do not always report their data to SEC and some only report results once per hour. The result is not all ionosondes contribute at each time period. For this particular 0830 UT case, only about half of the ionosondes contributed to the image. There are also several occultations represented in this particular example. Their influence upon the resulting analyzed electron density field is clearly visible in the lower left panel of Figure 2. Figure 2 also demonstrates how IDA3D treats occultation data. This particular case is for a PICOSat occultation over India at 0830 UT. The top left panel shows the original occultation data with a continuous line and the final IDA3D analysis fit to the data with a colored line. The top
right panel shows the geometry of the occultation. The curve represents the horizontal region of the ionosphere of the occultation TEC sampled. The triangle corresponds to the zero degree position of the lower panel plots and the pluses are marked at every 5° degrees of great circle distance. The lower left (right) panel shows a 2-D slice of analyzed (model) electron density, with the horizontal axis corresponding to the path given in the upper right panel and vertical axis giving altitude. The result clearly demonstrates a good fit to the data and shows that the IDA3D analysis provides corrections over the entire region affected by the data and not just a single profile.

[47] Figure 3 shows a comparison between the global analysis produced by IDA3D and the prediction from the background model PIM for 0400 UT. The upper panel shows the vertical TEC for the IDA3D analysis. The middle panel shows the vertical TEC obtained from PIM. The lower panel shows the percent difference between IDA3D and PIM. As expected, the largest percent differences occur in high data density regions, notably North America and Europe. The large ~25 percent variations below South America near Antarctica are due to GPS TEC data in the region, while the ~20 percent variations southeast of Australia are due to both GPS and occultation data.

[48] To demonstrate the three-dimensional aspects of IDA3D, latitude versus altitude slices of electron density are shown in Figure 4. The longitude of the slice is 310°E, which passes through the Greenland tomography array. The latitude ranges from 30 to 90 degrees geographic latitude. There is a great deal more smaller scale structuring in the IDA3D result during the night times and at high latitude where the tomography array is located.

[49] At ~0900 UT on 12 December, the $Kp$ increased from 2 to 5 and then retreated to a value of 3 at 1200 UT where the $Kp$ remained for the rest of the day. IDA3D was able to correctly track changes in the electron density that accompany the change in the magnetic activity level. Figure 5 shows four slices of electron density versus latitude and height. The slices are taken along a geomagnetic longitude of 31° along the Greenland tomography array axis. The slices are from 0800 to 0845 UT in 15 min increments. There is a definite evolution of the high-latitude auroral boundary in the 0845 UT image, as well as a definite increase in electron density in the low-latitude portion of the image.

[50] The residual chi-square fit to the data is $\chi^2 = \frac{1}{2} \left[ (\vec{y} - \vec{H} \vec{x}_a) R^{-1} (\vec{y} - \vec{H} \vec{x}_a) \right]$, where data outliers have been removed by the “buddy check” method described by Daley [1991]. Figure 6 presents histograms of the residual chi-

---

**Figure 1.** Example data coverage plot for 0830 UT. Dots are 350 km intercept points for GPS TEC, squares represent beacon tomography receivers data, triangles represent ionosondes, and the lines represent GPS occultations.
square. The resulting solid line histogram shows the residual fit to the data for the same four time periods as Figure 5. For reference, the dashed line shows the fit of the original model values to the data. In the upper left of each panel is give the residual mean and standard deviation for the IDA3D analysis and model values respectively. A “good residual fit” will have a standard deviation of ~1 and a mean of zero. As is seen from the figure, IDA3D produces a good fit for all four times. It is interesting to note the long negative tail in the model fit to the data, which demonstrates the non-Gaussian nature of the model.

6. Discussion
[51] This paper has described a new application of 3DVAR for globally specifying ionospheric electron density from data. The method presented here is closely based on the principles and methods widely used in the atmospheric

Figure 3. Comparison of Vertical TEC constructed by IDA3D (top) and PIM (middle) for 0400 UT, 12 December 2001. The lower panel gives the percent difference between IDA3D and PIM.
Figure 4. Vertical slice of electron density from IDA3D (left) and PIM (right) at a longitude of 310°E.
Figure 5. Four vertical slices of electron density from 0800 to 0900 UT. The slices are taken along a geomagnetic longitude of 31°E that is close to the longitude of the tomography array in Greenland.

Figure 6. Four plots showing the residual fit to the data for the four times presented in Figure 5. The residual error (solid line, $\bar{y} - H\bar{x}_b$) has been normalized by the data error input to the algorithm. The data innovation ($\tilde{y}^i - H\tilde{x}_b$) is shown with the dashed line.
data assimilation community. The analysis of electron density proceeds in a manner similar to traditional weather mapping. Starting from a background ionospheric specification (usually based upon an empirical model), the specification is modified to optimally agree with the available observations. This process produces specifications that capture the smaller-scale space weather features that are observed but not included in the background electron density field. Thus the resulting analyzed electron density field consists of the background model in regions where there is no influence of data and additional data-driven space weather features in data-rich regions.

The mathematical solution for the analyzed density field at any grid point is a weighted sum of background model and data. The weighting matrix is derived from the data and model error covariance matrices (equation (6)). Thus the analyzed electron density field only depends on measured data, the a priori information contained the initial model, and the error covariances. In principle, both the data and model error covariances could be obtained from knowledge of the underlying physics, and no ad hoc a priori constraints are required. This differs greatly from prior tomographic methods of inverting TEC data that rely on nonphysical smoothing constraints to obtain a solution.

An important aspect of IDA3D is its ability to use a number of different types of data. IDA3D ingests three types of TEC data: ground GPS TEC, ground CIT TEC, and satellite GPS occultation TEC. IDA3D also ingests peak density measurements from ionosondes and in situ measurements from satellites such as DMSP. IDA3D is able to accept all these data simultaneously, along with the various data instrumental and representational errors. IDA3D has been tested globally for various different geophysical conditions on a wide variety of data sets. For the data sets tested by IDA3D, the initial quality control initiated by the collecting agency is not always known, the data can vary significantly, and bad or poorly measured data can slip through the quality control. IDA3D has been able to produce a self-consistent, physically realistic analysis for all the test data sets regardless of the data quality and initial quality control. To demonstrate this, we have presented results for an entire day with the 15 min sampling time above. The capabilities of IDA3D are demonstrated in Figures 2, 3, 4, and 5 showing results for 12 December. Of particular interest are Figure 5 that demonstrates the ability of IDA3D to track even small changes in magnetic activity and Figure 2 that shows how occultation data is self-consistently ingested along with all the other data sources.

6.1. Future Improvements and Enhancements to IDA3D

There are a number of ways IDA3D can be enhanced. Additional data sources (such as satellite EUV/FUV limb scans) can be included and other background models, particularly physics-based predictive models, can be used as the initial background field. In addition, different model basis sets can be developed. Right now, IDA3D makes use of voxels, with the electron density constant within a voxel. However, other basis sets can be envisioned that may provide better accuracy. For example, a horizontal expansion in spherical harmonics [Howe et al., 1998] could be combined with a vertical basis of one-dimensional pixels, Bessel functions, or possibly nonlinear nine-parameter Chapman functions. While these research areas are potentially important enhancements to IDA3D, there are two basic areas of IDA3D that need to be addressed in the near future.

The first issue is data quality and quality control. IDA3D ingests a large amount of data from a variety of data sources. In the future, the number of data sources and the amount of data from each data source (i.e., the number of measurements from a given data type) will continue to increase. It is unrealistic to expect that each instrument data provider will have performed data quality and quality control at the level required for global assimilative analysis. IDA3D can be used to test quality control techniques, as well as how all the various data sources interact with each other in order to produce a self-consistent analysis. For example, IDA3D uses a buddy check procedure similar to Daley and Barker [2000] on the innovation vector and looks for outliers in the data to remove them. A better way to investigate and to understand the effects of the data is to sequentially add data sources and see how the results change. In particular, adding an occultation data set in the same region with DMSP and good ground GPS data can produce unrealistic profiles if the data and model error covariances have been mischaracterized. Finally, postanalysis of IDA3D electron densities often reveals the presence of bad data that has entered into the assimilation, either through unrealistically large or small densities. The second issue is the model covariances. One of the major physical limitations of IDA3D is the accuracy of the model error covariance. The covariance needs to be improved in two ways. First, a good parameterization of the model variance of electron density needs to be developed. The variance should be parameterized by conditions such as time of day, solar cycle, season, geomagnetic location, and magnetic activity level. Second, the spatial correlation model needs to be improved. Data sets need to be obtained and analyzed to determine if the Gaussian model adopted here is the appropriate functional form. In addition, empirical models of the spatial correlation lengths need to be developed with a parameterization similar that described for the variances.

Acknowledgments. Development of IDA3D is supported by the Office of Naval Research under grant N00014-97-1-0236. The instrumentation for the Greenland tomography array was built and deployed through the National Science Foundation under grant ATM-9813864. The Champ occultation data was provided courtesy of GFZ Postdam. The PICOSat occultation data was provided courtesy of Paul Straus, Aerospace Corp. Beacon CIT data from Qnanaq Greenland and the Alaska sites at Cordova, Delta Junction, and Gakona are provided by Northwest Research Associates (NWRA), while the Alaska sites at Poker Flat, Kaktovik, and Fort Yukon are provided by the Geophysical Institute, University of Alaska, Fairbanks. The ionosonde data was provided by the SPIDR web page (http://spidr.ngdc.noaa.gov) administrated by NOAA.

Arthur Richmond thanks Bruce Howe and another reviewer for their assistance in evaluating this paper.

References


G. S. Bust, T. W. Garner, and T. L. Gaussian II, Applied Research Laboratories, University of Texas at Austin, 10000 Burnett Street, Austin, TX 78759, USA. (gbust@arlut.utexas.edu)