Canonical correlation analysis of the combined solar wind and geomagnetic index data sets

Joseph E. Borovsky¹,²,³

¹Center for Space Plasma Physics, Space Science Institute, Boulder, Colorado, USA, ²AOSS, University of Michigan, Ann Arbor, Michigan, USA, ³Physics Department, Lancaster University, Lancaster, UK

Abstract
Canonical correlation analysis (CCA) will evaluate the degree of correlation between two multivariate data sets and will uncover patterns of correlation between the two data sets. Here CCA is applied to the multivariate solar wind data set and the multivariate geomagnetic index data set. CCA creates a new set of solar wind variables and a new set of Earth variables. The first of the new solar wind variables can be used as a solar wind driver function for the magnetosphere; the conjugate Earth variable can be used as an Earth vector to describe global geomagnetic activity in the magnetosphere-ionosphere system. The CCA-generated driver functions are found to be superior in accuracy to other driver functions in the literature. CCA of the combined data sets provides information about (1) differences in the solar wind driving of high-latitude geomagnetic indices (AE, AU, AL, and polar cap index) versus magnetospheric-convection geomagnetic indices (Kp and midnight boundary index), (2) the properties of electric field-based driver functions versus reconnection-based driver functions, and (3) improvements to solar wind/magnetosphere correlations produced by time averaging the solar wind clock angle. The CCA process tends to focus on the magnetospheric-convective indices over other indices: this may indicate that there is more predictable variance in the global-convective indices than in the others.

1. Introduction
The solar wind-driven magnetosphere is a highly connected system [Vassiliadis, 2006; Valdivia et al., 2013; Borovsky, 2014]: causes have multiple effects, and outcomes have multiple causes. One approach to a systems science of the magnetosphere is to perform multivariate analysis of the driving of the magnetospheric system by the solar wind. Canonical correlation analysis is an ideal tool when causes and effects cannot be described or measured by single variables [Nimon et al., 2010; Hair et al., 2010].

As the properties of the solar wind at the Earth vary, activity in the Earth’s magnetosphere-ionosphere system varies. For some decades the interaction of the solar wind with the magnetosphere coupling has been statistically studied. Typically, a time series of one function of measured solar wind variables is cross correlated with a time series of values of one of the geomagnetic indices. The magnitude of the correlation coefficient is interpreted as a measure of the accuracy of the function of solar wind variables to describe the physics of the driving of the Earth’s magnetosphere by the solar wind.

In contrast, in this report the solar wind data set will be collectively compared with the geomagnetic index data set. Treating the multivariate geomagnetic index data set collectively is like analyzing the stock market rather than analyzing the behavior of one stock. With the “first canonical correlation,” the canonical correlation analysis process can be used to develop canonical solar wind driver functions that can predict the “stock market” of geomagnetic activity. (The resulting canonical driver functions will also be very good at predicting individual geomagnetic indices.) With “second” and “third canonical correlations,” the canonical correlation analysis process can be used to study what drives behavior patterns within the stock market of geomagnetic indices.

Via the correlation method, two approaches have been followed to obtain solar wind driver functions for the magnetosphere [cf. Borovsky and Birn, 2014]. First, driver functions based on the functional form of the solar wind motional electric field (i.e., based on \( v_{sw} \) and \( B_z \)) have been written and tested [Wygant et al., 1983; Reff and Luhmann, 1986; Newell et al., 2007]. The functional forms of these electric field driver functions include (see Table 1 for symbol definitions) \( v_{sw}B_z \) [e.g., Hardy et al., 1981], \( v_{sw}B_z \) [e.g., Holzer and Slavin, 1982], \( v_{sw}B_z \sin^2(\rho_{clock}/2) \) [Kan and Lee, 1979], \( v_{sw}B_z n^{1/2}\sin^4(\rho_{clock}/2)/(1 + 5 \times 10^{-5}|B_z|^2) \) [Tenford and Østgaard, 2013];
and $v_{sw}^{4/3}B_z^{2/3}\sin^{2/3}(\theta_{\text{clock}}/2)$ [Newell et al., 2007]. A second approach [Borovsky, 2008] has been used to derive driver functions from the Cassak-Shay equation [Cassak and Shay, 2007; Birn et al., 2012] for asymmetric reconnection: two such functions are $R_{\text{quick}}$ [Borovsky and Birn, 2014] and $R_2$ [Borovsky, 2013a]. It has also been argued that a viscous interaction between the solar wind and the magnetosphere adds to the solar wind driving [cf. Vasyliunas et al., 1982; Borovsky and Funsten, 2003; Borovsky, 2006, 2013b; Newell et al., 2008].

A few years ago the author and Joachim Birn accidentally arrived at a new class of solar wind driver functions for the magnetosphere. Using a computer code specifically written to generate, test (by correlation), and evolve (by mutations) solar wind driver functions for the magnetosphere, the function $v_{sw} + 56B_z$ was produced. (The coefficient 56 will vary from data set to data set.) The solar wind function $v_{sw} + 56B_z$ produces a significantly higher correlation with geomagnetic indices than any of the electric field functions or the reconnection functions. The function $v_{sw} + 56B_z$ has no obvious physical interpretation. That it does so well at predicting geomagnetic activity poses a dilemma since we use improved correlations as our test of the correctness of physics-based solar wind driver functions. Specifically,
The polar cap index PCI is a direct measure of the strength of transpolar ionospheric current. The various geomagnetic indices measure various properties of the magnetosphere-ionosphere system. Physics-based variables are the measured properties of the solar wind plasma: $n$, $T$, $V_{sw}$, $V_P$, $V_P$, $n$, $T$, $V_{sw}$, $V_P$, $B_n$, $B_P$, the alpha-particle number density, the heavy ion charge-state density ratios, amplitudes of fluctuations, etc. Here what is meant by physics-based variables are the quantities that our physical analysis indicates the Earth might care about. Such examples of physics-based variables are ram pressure $nV_{sw}^2$ (which determines the magnetic field strength at the nose of the magnetosphere, which affects dayside reconnection rate), $\theta_{\text{clock}}$ (which affects the magnetic clock angle across the dayside magnetopause, which affects the component rate of dayside reconnection), $n$ (which affects the mass density of the magnetosheath plasma, which affects the dayside reconnection rate), the Alfvén Mach number $M_A = B_{\text{mag}}/(4\pi n e n)^{1/2}$ (which affects the strength and position of the bow shock, which determines the plasma-beta of the magnetosheath and the flow pattern of the magnetosheath, both of which alter dayside reconnection rates), etc. Note that the solar wind number density $n$ can be considered to be both a primitive variable and a physics-based variable. In the example of section 3 (Table 2) the physical solar wind variables $\log(nV_{sw}^2)$, $\log(n)$, $\log(M_A)$, and $\theta_{\text{clock}}$ were used as input; these variables contain the same information as the more primitive variables $\log(n)$, $\log(n)$, $\log(B_{\text{mag}})$, and $\theta_{\text{clock}}$.

Likewise, the geomagnetic index data set with variables polar cap index (PCI), $AU$, $AL$, $Kp$, midnight boundary index (MBI), $Dst$, … is a multivariate data set with strong intercorrelations among the variables. The various geomagnetic indices measure various properties of the magnetosphere-ionosphere system. The polar cap index PCI is a direct measure of the strength of transpolar ionospheric current [Troschichev et al., 1988; Stauning, 2013] and is related to polar cap antisunward convection [Stauning, 2013] and cross-polar cap potential [Ridley and Kihn, 2004]; PCI shows reactions to the occurrence of substorms, although the strength of the reaction varies with season [Janzhura et al., 2007]. The $AU$ index is a measure of the maximum eastward auroral electrojet in the ionosphere [Davis and Sugiuara, 1966] which in active times typically occurs in the dusk region [Allen and Kroehl, 1975]; the $AU$ index is associated with the DP-2

### Table 2. For the Physical-Variable CCA Example of Section 3, the Coefficients (Weights) of the First Canonical Variables $E_{(1)}$ and $S_{(1)}$ Are Collected Into the First and Third Columns, and the Correlation Coefficients (Loadings) Between the Individual Input Variables and $E_{(1)}$ and $S_{(1)}$ Are Collected Into the Second and Fourth Columns

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient in $E_{(1)}$</th>
<th>Correlation With $E_{(1)}$</th>
<th>Coefficient in $S_{(1)}$</th>
<th>Correlation With $S_{(1)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earth variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{PCI}^*$</td>
<td>0.121</td>
<td>80.9%</td>
<td>68.9%</td>
<td></td>
</tr>
<tr>
<td>$\log(AE_{(1)})^*$</td>
<td>0.331</td>
<td>91.9%</td>
<td>78.2%</td>
<td></td>
</tr>
<tr>
<td>$Kp_{(1)}^*$</td>
<td>0.402</td>
<td>93.6%</td>
<td>79.7%</td>
<td></td>
</tr>
<tr>
<td>$-\text{MBI}_{(1)}^*$</td>
<td>0.238</td>
<td>92.9%</td>
<td>79.0%</td>
<td></td>
</tr>
<tr>
<td>Solar wind variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\log(nV_{sw}^2)^*$</td>
<td></td>
<td>38.5%</td>
<td>1.196</td>
<td></td>
</tr>
<tr>
<td>$\log(n)$</td>
<td>-0.7%</td>
<td>-0.784</td>
<td>-1.1%</td>
<td></td>
</tr>
<tr>
<td>$\log(M_A)$</td>
<td>-20.7%</td>
<td>-0.388</td>
<td>-24.4%</td>
<td></td>
</tr>
<tr>
<td>$\theta_{\text{clock}}^*$</td>
<td></td>
<td>50.3%</td>
<td>0.602</td>
<td></td>
</tr>
</tbody>
</table>

If a change to a driver function results in an improved correlation with geomagnetic activity, is it because the physics of the changed driver function is now more correct or is it because the changed driver function has a mathematical advantage? If it is the latter, how can we test the correctness of the physics of a driver function?
current system [Tomita et al., 2011] and is sometimes considered as a measure of dayside high-latitude field-aligned currents, or at least it is associated with field-aligned currents that have a short Alfvén wave transit time [Goertz et al., 1993]. (For the DP-1 and DP-2 current systems, see Clauer and Kamide [1985].) The AL index is a measure of the maximum westward auroral electrojet in the ionosphere [Davis and Sugiyama, 1966] which during active times typically occurs in the postmidnight region [Allen and Kroehl, 1975]; AL is associated with both the DP-2 and DP-1 field-aligned current systems [Tomita et al., 2011], and so AL reacts to the occurrence of substorms [Prichard et al., 1996; Troshichev et al., 2012], and AL is sometimes considered as a measure of nightside high-latitude field-aligned currents [Goertz et al., 1993]. The AE index is the sum of the magnitudes of AU and AL and has been considered to be a measure of the total auroral electrojet current although that has been criticized [Kamide and Rostoker, 2004]; AE has been used as a proxy for the total amount of Joule dissipation in the high-latitude ionosphere [Baumjohann, 1986]. The midnight boundary index MBI is a measure of the latitude of the equatorward boundary of the diffuse aurora at local midnight [Gusenhoven et al., 1983], which magnetically maps to the radial location of the inner edge of the electron plasma sheet, which is a direct measure of the strength of global magnetospheric convection [Elphic et al., 1999; Denton et al., 2005]. The Kp index is sensitive to the latitude of currents at the inner edge of the electron plasma sheet [Thomsen, 2004], which is directly related to the strength of global plasma convection in the magnetosphere [Elphic et al., 1999; Denton et al., 2005]. The ram pressure-correctedDst index $DSt^*$ is a measure of the plasma diamagnetic current in the inner magnetosphere [Dessler and Parker, 1959] produced by ions with orbits trapped in the dipolar magnetosphere [Scopke, 1966] and by plasma flowing past the dipole from the nightside to the dayside [Liemohn et al., 2001]; $DSt^*$ is also sensitive to the strength and downtail location of cross-tail currents [Turner et al., 2000; Alexeev and Feldstein, 2001; Ohtani et al., 2001; Borovsky and Denton, 2010]. Since $DSt^*$ largely depends on the properties of the plasma in the inner magnetosphere [Thomsen et al., 1998; Moore et al., 2005], plasma transport timescales produce longer time lags in the response of $DSt^*$ to the solar wind [Smith et al., 1999].

In this report canonical correlation analysis is utilized to mathematically look for patterns and correlations between the solar wind and geomagnetic index multivariate data sets. Physical interpretations of the mathematical results are made where possible.

This report is organized as follows. In section 2 an overview is given of the canonical correlation analysis technique. In section 3 an example canonical correlation for solar wind/magnetosphere coupling is discussed. In section 4 the first canonical correlation, in general, is examined, and in section 5 the second canonical correlation is interpreted. In section 6 canonical correlation analysis is used to generate and examine solar wind driver functions for the magnetosphere. The findings are summarized in section 7, and future research is discussed in section 8. For the nonspace physics specialists, the definitions of the symbols used in this report are listed in Table 1.

2. Canonical Correlation Analysis

Canonical correlation analysis (CCA) will evaluate the degree of correlation between two multivariate data sets, and CCA finds patterns and correlations between two multivariate data sets [cf. Muller, 1982; Johnson and Wichern, 2007; Gatignon, 2010; Nimon et al., 2010]. Two such multivariate data sets are the solar wind data set and the geomagnetic index data set. Both data sets are composed of variables that are functions of time. When applied, CCA creates new composite solar wind variables $S_{11}, S_{22}, S_{33}, \ldots$ that are linear combinations of the input solar wind variables, and CCA creates new composite Earth variables $E_{11}, E_{22}, E_{33}, \ldots$ that are linear combinations of the input geomagnetic indices. Like the original input variables, the new composite variables $S_{11}, S_{22}, \ldots$ and $E_{11}, E_{22}, \ldots$ are functions of time. The pair $S_{11}$ and $E_{11}$ are the first canonical variables of the combined data set and the correlation coefficient $r_{11}$ between them is the “first canonical correlation” between the two data sets. The variables $S_{1k}$ and $E_{1k}$ are the “kth canonical variables” and the correlation coefficient $r_{kk}$ between them is the “kth canonical correlation” between the two data sets. The variable $S_{1k}$ is correlated with itself and with $E_{1k}$ and is uncorrelated with all other $S_{ij}$ and $E_{ij}$ unless $i = k$. Likewise, variable $E_{1k}$ is correlated with itself and with $S_{1k}$ and is uncorrelated with all other $S_{ij}$ and $E_{ij}$ unless $i = k$. The new composite solar wind $S$ variables are linear combinations of the input solar wind variables “$V$” and the new composite Earth $E$ variables are linear combinations of the input geomagnetic indices “$G$”. 

---

**References**

- Tomita et al., 2011
- Goertz et al., 1993
- Liemohn et al., 1986
- Dessler and Parker, 1959
- Turner et al., 2000
- Alexeev and Feldstein, 2001
- Ohtani et al., 2001
- Borovsky and Denton, 2010
- Thomsen et al., 1998
- Moore et al., 2005
- Smith et al., 1999
- Clauer and Kamide, 1985
- Baumjohann, 1986
- Muller, 1982
- Johnson and Wichern, 2007
- Gatignon, 2010
- Nimon et al., 2010
- Muller, 1982
The $k$th $S_{(k)}$ and $E_{(k)}$ time series variables can be written as follows [cf. Johnson and Wichern, 2007; Gatignon, 2010]:

\[
S_{(k)}(t) = a_{(k)} \cdot V(t) \quad (1a)
\]

\[
E_{(k)}(t) = b_{(k)} \cdot G(t) \quad (1b)
\]

where at every value of time $V$ is the vector of input solar wind variables, where at every value of time $G$ is the vector of input geomagnetic indices, where $a_{(k)}$ is the vector of coefficients that transforms $V$ into $S_{(k)}$, and where $b_{(k)}$ is the vector of coefficients that transforms $G$ into $E_{(k)}$.

The information needed to obtain the canonical variables $S$ and $E$ is contained in the four covariance matrices between the original input solar wind variable $V$ and the original input geomagnetic variable $G$: a matrix of correlation coefficients within the solar wind data set, a matrix of correlation coefficients within the geomagnetic index data set, and two matrices of correlation coefficients between the variables of the solar wind and the geomagnetic indices. Denote $\Sigma_{vv}$ as the correlation matrix among the solar wind variables $V$; if there are $N$ solar wind variables, then $\Sigma_{vv}$ will be an $N \times N$ matrix of the cross correlations between the solar wind variables. Each element of the $\Sigma_{vv}$ matrix is a Pearson linear correlation coefficient between two solar wind variables obtained by cross correlating the two times series of values of the two variables. For example, if the three variables $V_{sw}$, $n$, and $B_{mag}$ are used for the solar wind data set, then the $3 \times 3$ matrix elements of $\Sigma_{vv}$ are the correlation coefficients between the three variables. Using "1" for $V_{sw}$, "2" for $n$, and "3" for $B_{mag}$ the nine elements of the $\Sigma_{vv}$ matrix are as follows: the 1,1 element ($V_{sw} \leftrightarrow V_{sw}$) is 1.000, the 1,2 element ($V_{sw} \leftrightarrow n$) is $-0.358$, the 1,3 element ($V_{sw} \leftrightarrow B_{mag}$) is 0.178, the 2,2 element ($n \leftrightarrow n$) is 1.000, the 2,3 element ($n \leftrightarrow B_{mag}$) is 0.285, and the 3,3 element ($B_{mag} \leftrightarrow B_{mag}$) is 1.000. The matrix is symmetric, so the $ij$ element equals the $ji$ element. Denote $\Sigma_{gg}$ as the correlation matrix among the geomagnetic indices; if there are $M$ geomagnetic indices, then $\Sigma_{gg}$ will be an $M \times M$ matrix of the cross correlations between the various geomagnetic indices. Again, $\Sigma_{gg}$ is a symmetric matrix with diagonal elements of unity. Denote $\Sigma_{vg}$ and $\Sigma_{gv}$ as the correlation matrices between the solar wind variables and the geomagnetic indices; $\Sigma_{vg}$ will be an $N \times M$ matrix, and $\Sigma_{gv}$ will be an $M \times N$ matrix with $\Sigma_{vg}$ being the transpose of $\Sigma_{gv}$. The mathematical derivation of the CCA processes is based on maximizing the Pearson linear correlation between the pairs of composite variables $S_{(k)} \leftrightarrow E_{(k)}$. This maximizing condition leads, after substantial matrix algebra [cf. Johnson and Wichern, 2007; Gatignon, 2010], to two eigenvalue problems. The eigenvectors and square roots of the eigenvalues of the matrix $\Sigma_{vv}^{-1} \Sigma_{vg} \Sigma_{gg}$ yield the coefficient vectors $a_{(k)}$ and the canonical correlations $S_{(k)}$ and the eigenvectors and square roots of the eigenvalues of the matrix $\Sigma_{gg}^{-1} \Sigma_{gg} \Sigma_{gv} \Sigma_{vv}^{-1} \Sigma_{vg}$ yield the coefficient vectors $b_{(k)}$ and the same canonical correlations $E_{(k)}$. The largest eigenvalue (largest squared correlation coefficient $r^2$) corresponds to the first canonical variables $S_{(1)}$ and $E_{(1)}$; the second largest eigenvalue corresponds to the second canonical variables $S_{(2)}$ and $E_{(2)}$, etc.

A geometrical interpretation of the matrices $\Sigma_{vv}^{-1} \Sigma_{vg} \Sigma_{gg}^{-1} \Sigma_{gv}$ and $\Sigma_{gg}^{-1} \Sigma_{gg} \Sigma_{gv} \Sigma_{vv}^{-1} \Sigma_{vg}$ [Johnson and Wichern, 2007] is that the correlation matrix $\Sigma_{vv}$ is used to create a principal component set of solar wind variables and that the correlation matrix $\Sigma_{gg}$ is used to create a principal component set of Earth variables. The information in the cross-correlation matrices $\Sigma_{vg}$, and $\Sigma_{gv}$ is then used to separately rotate the principal-coordinate representations of the solar wind data set and the geomagnetic index data set into projections onto each other to yield the maximized cross correlations.

Although not required, CCA will perform best with input variables that are Gaussian distributed [Hair et al., 2010]. For positive-definite variables that have very skew distributions (such as $n$, $V_{sw}^2$, or $AE$) using the logarithm of the variable produces higher correlation coefficients in the CCA process. However, in general, the solar wind variables and geomagnetic index variables are not Gaussian distributed.

For the input solar wind variables $V$ and the input geomagnetic indices $G$ we will normalize (standardize) the variables by subtracting off the mean value of the variable and dividing by the standard deviation of the variable. This also makes the variable dimensionless. For example, for use in CCA the solar wind number density $n$ will be transformed to $n^* = (n - \langle n \rangle)/\sigma(n)$ where $\langle n \rangle$ is the mean value of $n$ and $\sigma(n)$ is the standard deviation of $n$. If the natural (base e) logarithm of $n$ is used, then log$(n)$ will be standardized into $\log(n)^* = [\log(n) - \langle \log(n) \rangle]/\sigma(\log(n))$. When the input variables are all dimensionless and normalized with zero mean and with standard deviations of unity, then all of the canonical variables $S_{(k)}$ and $E_{(k)}$ will also be dimensionless and with zero mean and standard deviations of unity.
To perform a CCA analysis of the solar wind and geomagnetic index data sets, the following sequence of tasks are performed: (1) Choose the input variables from the solar wind and choose the input variables from the geomagnetic indices, (2) eliminate missing data, (3) standardize the input variables by removing their mean values and dividing by their standard deviations, (4) calculate all of the Pearson correlation coefficients between all of input variables, (5) use the correlation coefficients to create the four matrices $\Sigma_{vv}$, $\Sigma_{gg}$, $\Sigma_{vg}$, and $\Sigma_{gv}$ (6) form the two combined matrices $\Sigma_{vv}^{-1}\Sigma_{vg}^{-1}\Sigma_{gv}$ and $\Sigma_{gg}^{-1}\Sigma_{vg}^{-1}\Sigma_{gv}$ and (7) calculate the eigenvectors and eigenvalues of the two combined matrices.

Note that among the variables of the solar wind data set there are many intercorrelations and that among the variables of the geomagnetic index data set there are many intercorrelations. In CCA, as in any statistical analysis, multicollinearity makes physical interpretation of cause and effect difficult and multicollinearity makes discerning the impact of any one variable difficult [Lambert and Durand, 1975; Hair et al., 2010; Nimon et al., 2010]. However, for dealing with multicollinearity, CCA may provide an advantage because patterns that contain differences of (subtractions) correlated variables appear in the second, third, and beyond canonical correlations. These differenced variables may reveal the importance or influence of the variances of the variables that are not intercorrelated.

By examining the magnitude of the coefficient of an input variable going into a canonical composite variable and comparing it with the magnitude of the correlation coefficient between the input variable and the canonical composite variable [cf. Hair et al., 2010; Nimon et al., 2010], one can obtain information about (1) the importance of that input variable to deriving that composite variable, (2) the relative contribution that the input variable makes to the composite variable, (3) whether the input variable is acting as a suppressor in the composite variable (acting to cancel out the irrelevant variance of other input variables in order to improve the overall correlations [cf. Conger, 1974; Tzelgov and Henik, 1991]), and (4) whether the input variable’s effect in the composite variable is also explained by other input variables [cf. Robins, 1989; Frank, 2000].

Like simpler statistical methods, what you can interpret depends on the input variables that you use. With canonical correlation analysis, the number of simultaneous input variables is not limited, so the choice of input variables can be quite complicated. Different sets of input variables yield different results and can be used to study different problems. In this report a limited number of cases will be explored. Other cases have been explored in subsequent reports [Borovsky and Denton, 2014; J. E. Borovsky, The efficacy of solar-wind/magnetosphere coupling through the solar cycle examined with an accurate driver function: No dependence on the phase of the solar cycle, submitted to Journal of Geophysical Research, 2014].

### 3. An Example Canonical Correlation

In this example, for the solar wind data set the four physical variables $\log(n_{sw}^2)^c$, $\log(n)^c$, $\log(M_A)^c$, and $\theta_{\text{clock}}^c$ are taken (where $\log = \log_e = \ln$ is the natural logarithm), and for the Earth data set the four geomagnetic indices $\text{PC1}_v^c$, $\log(\text{AE}_1)^c$, $K_p^c$, and $\text{MB}_1^c$ are taken, where the $^c$ is a reminder that standardized values are used. On the geomagnetic indices, the subscript number is the number of hours of time lag on the index relative to the solar wind; e.g., $\text{AE}_1$ is the hourly value of $\text{AE}$ taken 1 h after the hourly value of the measured solar wind. These optimal lag times for the various geomagnetic indices have been established in earlier studies [cf. Borovsky, 2013b] and are adopted here. Each variable has 102,900 h of measurements: considerably more hours of data are available in the OMNI2 solar wind data set [King and Papitashvili, 2005], but the restriction to 102,900 h comes from two requirements imposed in this report (1) that measurements of the Earth’s midnight boundary index $\text{MB}_1$ [Madden and Gussenhoven, 1990] must be included and (2) that the solar wind alpha-to-proton density ratio must be included. An hour of data is not included unless off solar wind variables and all geomagnetic indices are available. The 102,900 h of data are from the years 1983 to 2006.

For this four-variable solar wind data set and four-variable geomagnetic index data set as input into CCA, the first canonical variables are

\begin{align}
S_{(1)} & = 1.196 \ \log(n_{sw}^2)^c - 0.784 \ \log(n)^c - 0.388 \ \log(M_A)^c + 0.602 \ \theta_{\text{clock}}^c \\
E_{(1)} & = 0.121 \ \text{PC1}_v^c + 0.331 \ \log(\text{AE}_1)^c + 0.402 \ K_p^c - 0.238 \ \text{MB}_1^c
\end{align}

These coefficients are listed in the first and third columns of Table 2. For this first pair of canonical variables, the first canonical correlation between $S_{(1)}$ and $E_{(1)}$ is $r_{(1)} = 0.851$. For the 102,900 data points, $E_{(1)}$ is plotted as
the coefficient (weight) of \( \log(n)^* \) is nonzero \((-0.388)\), whereas its correlation coefficient (loading) with \( E_{(1)} \) is essentially zero \((-0.7\%)\). This is an example of a variable \( \log(n)^* \) playing a suppressor role in the composite variable \( S_{(1)} \) \( [\text{Conger, 1974; Tzelgov and Henik, 1991}] \): its role is to cancel off (suppress) some variance in another input variable, where that canceled variance is uncorrelated with \( E_{(1)} \).

\( S_{(1)} \) can be used as a general driver function that does a very good job of predicting the variance of \( E_{(1)} \) and also does a fairly good job of predicting the individual geomagnetic indices. Better driver functions generated by CCA will be seen in section 6.

In expression (2a), replacing the standardized * variables with their nonnormalized forms (i.e., \( \log(n)^* = (\log(n) - \langle \log(n) \rangle) / \sigma(\log(n)) \)) and using \( a \log(b) + c \log(d) = \log(b^a d^c) \), the first canonical solar wind variable \( S_{(1)} \) can be written as

\[
S_{(1)}/2.055 = \log\left(3.37 \times 10^{-6} \text{nsvw}^2 n^{-0.561} M_A^{-0.466} \exp(\theta_{\text{clock}}/146) \right). \tag{3}
\]

Inside the logarithm of expression (3), the form of the variables of \( S_{(1)} \) has a resemblance to \( R_{\text{quick}} \) where \( [\text{cf. Borovsky and Birn, 2014}] \)

\[
R_{\text{quick}} = \left(\text{nsvw}^2 / n^{1/2}\right) \sin^2(\theta_{\text{clock}}/2) M_A^{-1.35} \left[1 + 680 M_A^{-3.30}\right]^{-1/4}. \tag{4}
\]

inside the logarithm of expression (3) the power of \( \text{nsvw}^2 \) is unity, as it is in \( R_{\text{quick}} \) and inside the logarithm the power of \( n \) is \(-0.561\) which is close to the power of \( n \) of \(-0.5 \) in \( R_{\text{quick}} \). In comparing expression (3) with expression (4), the Mach number dependence of \( S_{(1)} \) is shallower than the Mach number dependence of \( R_{\text{quick}} \) which varies as \( M_A^{-0.52} \) at low Mach number and as \( M_A^{-1.35} \) at high Mach number. This difference in Mach number dependence between \( S_{(1)} \) and \( R_{\text{quick}} \) might yield information about the Mach number dependence of the length of the dayside reconnection X-line, which is not in the derivation of \( R_{\text{quick}} \) but which must be important for solar wind/magnetosphere coupling. A theoretical argument for the length of the reconnection X-line on the dayside magnetosphere as a function of the Mach number was presented in Borovsky [2013b] based on the flow-shear pattern at the dayside magnetopause as a function of \( M_A \); that argument predicted that the X-line would be shorter as \( M_A \) increased, which would produce a total reconnection rate that has a steeper \( M_A \) dependence than that of expression (4), opposite to the trend of \( S_{(1)} \).

In Figure 2 the function \( \exp(\theta_{\text{clock}}/91)/87 \) of expression (3) is compared with \( \sin^2(\theta_{\text{clock}}/2) \) of expression (4).
As can be seen in the fourth column of Table 3, the Pearson linear correlation coefficient between the two variables and very different indices, it is common that the correlation between these two functions is $r_{corr} = 0.918$. The 300-point running average (blue) indicates a linearity between the two functions.

### 4. General Results: First Canonical Correlations

The first canonical correlation produces the linear combinations of input variables that yield the highest correlation between the two multivariate data sets. It is not yet known if the canonical correlation coefficients are higher if physical quantities such as $n v_{sw}^2$ and $M_A$ are used as inputs instead of primitive solar wind variables such as $v_{sw}$ and $B_{mag}$. For comparison with the example of section 3, if the primitive solar wind variables $\log(v_{sw}), \log(B_{mag}), \log(n)$, and $\theta_{clock}$ are used with the same geomagnetic index set of variables ($PCI_o, \log(AE_1), Kp_1$, and $MBI_1$), the CCA procedure yields the first canonical variables

\[
S_{(1)} = 0.704 \log(v_{sw})^* + 0.386 \log(B_{mag})^* + 0.302 \log(n)^* + 0.601 \theta_{clock}^* \\
E_{(1)} = 0.120 PCI_o^* + 0.331 \log(AE_1)^* + 0.403 Kp_1^* - 0.239 MBI_1^* 
\]

with a first canonical correlation of $r_{(1)} = 0.851$. This value of $r_{(1)}$ is the same as that of the example of section 3 using the physical variables. The coefficients of $S_{(1)}$ and $E_{(1)}$ are displayed in the third and first columns of Table 3, respectively. The coefficients of $E_{(1)}$ in expression (5b) are essentially the same as those of expression (2b). It is also the case that, to within 1 percentage point, $S_{(1)}$ of expression (2a) and $S_{(1)}$ of expression (5a) are equal. Using $M_A = v_{sw}/v_A = v_{sw}^{1/2}/21.8 B_{mag}$ (where $v_{sw}$ is in units of km/s, $n$ is in units of cm$^{-3}$, and $B_{mag}$ is in units of nT) and writing all variables in nonstandardized forms, expressions (2a) and (5a) both yield

\[
S_{(1)} = 3.19 \log(v_{sw}) + 0.445 \log(n) + 0.916 \log(B_{mag}) \\
+ 0.0138 \theta_{clock} - 23.0 
\]

Note that if CCA is performed with the solar wind input variables not inside of logarithms, then the $nv_{sw}^2$, $n$, $M_A$ and $\theta_{clock}$ data input case and the $v_{sw}$, $B_{mag}$, $n$, and $\theta_{clock}$ data input case yield very different first canonical variables and very different first canonical correlation coefficients.

As can be seen in the fourth column of Table 3, the Pearson linear correlation coefficient between the function $S_{(1)}$ given by expression (5a) and the individual geomagnetic indices in expression (5b) is 68.9% for $PCI_o$, 78.2% for $\log(AE_1)$, 79.7% for $Kp_1$, and 79.1% for $MBI_1$. For the physical-variable case of Table 2 the fourth column correlation coefficients between $S_{(1)}$ and the geomagnetic indices are the same. In general, examining the correlation coefficients between the first canonical solar wind variables $S_{(1)}$ and the individual geomagnetic indices, it is common that the correlation between $S_{(1)}$ and global-convection indices ($Kp$ and $MBI$) is higher than the correlations between $S_{(1)}$ and the high-latitude indices ($AE$, $AU$, $AL$, and $PCI$). CCA yields functions that are better at predicting global-convection indices than high-latitude indices. The CCA procedure seems to indicate that there is more predictable variance in the global-convection geomagnetic indices than there is in the high-latitude indices. Note in Table 3, as was the case in Table 2, that $\log(n)^*$ is playing the role of a suppressor variable: its coefficient (weight) in $S_{(1)}$ is nonzero, but its correlations are near zero.
The coefficient of the first canonical correlation increases as the number of solar wind input variables increases. This could be because more information is being added to the analysis, or it could be because more free parameters are available for fitting. For the four variables of the CCA example in section 3 \( \log(n_{sw})^2 \), \( \log(n) \), \( \log(M_A) \), and \( \sin^2(\theta_{clock}/2) \) this is explored in Table 4 where a “cross” indicates the inclusion of a solar wind variable as input to CCA. The second-to-last column is the first canonical correlation coefficient \( r_{(1)} \), and the last column of Table 4 is the unaccounted for variance \( 1 - r_{(1)}^2 \) in percent. (The shared variance is \( r_{(1)}^2 \) and the unaccounted for variance is \( 1 - r_{(1)}^2 \)) For all cases in Table 4 seven geomagnetic indices \( (AU_1, -AL_1, \log(AE_1), PC_{ly}, KP_1, -MBI_1, \text{ and } -Dst^*_2) \) are used as input. There are several different formulas for creating the pressure-corrected \( Dst^* \) index from the \( Dst \) index [e.g., Su and Konradi, 1975; Burton et al., 1975; Feldstein et al., 1984; Pudovkin et al., 1985; Vassiliadis et al., 1999; O’Brien and McPherron, 2000]; the Borovsky and Denton [2010] formula \( Dst^* = Dst - bP_{ram}^{1/2} + c \) with \( b = 20.7 \) nT(nPa)\(^{-1/2} \) and \( c = 27.7 \) nT (which comes from a fit to \( Dst \) versus \( P_{ram} \) rather than a fit to \( dDst/dP_{ram} \)) will be used here. In the first four rows of Table 4 the first canonical correlation \( r_{(1)} \) is displayed for the individual core variables. Row 5 of Table 4 uses all four core variables: the resulting \( r_{(1)} \) value is 0.859. The rows below row 5 in Table 4 add other solar wind variables one at a time into the CCA with the core variables \( \log(n_{sw})^2 \), \( \log(n) \), \( \log(M_A) \), and \( \sin^2(\theta_{clock}/2) \). As can be seen for this set of core solar wind variables, some variables (e.g., \( B_2 \)) produce a noticeable improvement in the value of \( r_{(1)} \) and some variables (e.g., \( v_{sw} \) and \( B_{mag} \)) do not. The statistical significance of some of the small increments in \( r_{(1)} \) are questionable, and the variables that yield those small increments are of limited importance for the canonical correlation. Other variables (not shown in Table 4) make no change in the value of \( r_{(1)} \). In the final row of Table 4 an 11-variable solar wind data set is fed into CCA: for this 11-variable case the first canonical correlation is \( r_{(1)} = 0.894 \) with the corresponding first canonical Earth variable \( E_{(1)} \).

Borovsky [2013b] found that the removal of uncorrelated high-frequency noise from the solar wind data set improves solar wind/magnetosphere correlation coefficients [see also Newell et al., 2007], and of course, time averaging both the solar wind and the geomagnetic indices yields higher correlations [Kamide, 1983; Kamide et al., 1984; Vassiliadis et al., 1999; O’Brien and McPherron, 2000].

**Figure 3.** Hourly averaged values of the interior of the logarithm of expression (3) are plotted as a function of hourly averaged values of the reconnection driver function \( R_{quick} \). The Pearson linear correlation coefficient of the plotted points is 0.918.

**Table 3.** For the Primitive-Variable CCA Example of Section 4, the Coefficients (Weights) of the First Canonical Variables \( E_{(1)} \) and \( S_{(1)} \) are Collected into the First and Third Columns, and the Correlation Coefficients (Loadings) Between the Individual Input Variables and \( E_{(1)} \) and \( S_{(1)} \) are Collected into the Second and Fourth Columns.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient in ( E_{(1)} )</th>
<th>Correlation With ( E_{(1)} )</th>
<th>Coefficient in ( S_{(1)} )</th>
<th>Correlation With ( S_{(1)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earth variables</td>
<td>( PC_{ly}^* )</td>
<td>0.120</td>
<td>80.9%</td>
<td>68.9%</td>
</tr>
<tr>
<td></td>
<td>( \log(AE_1)^* )</td>
<td>0.331</td>
<td>91.8%</td>
<td>78.2%</td>
</tr>
<tr>
<td></td>
<td>( KP_1^* )</td>
<td>0.403</td>
<td>93.7%</td>
<td>79.7%</td>
</tr>
<tr>
<td></td>
<td>( -MBI_1^* )</td>
<td>0.239</td>
<td>92.9%</td>
<td>79.1%</td>
</tr>
<tr>
<td>Solar wind variables</td>
<td>( \log(v_{sw})^* )</td>
<td>52.4%</td>
<td>0.704</td>
<td>61.5%</td>
</tr>
<tr>
<td></td>
<td>( \log(B_{mag})^* )</td>
<td>47.5%</td>
<td>0.386</td>
<td>55.8%</td>
</tr>
<tr>
<td></td>
<td>( \log(n)^* )</td>
<td>–1.0%</td>
<td>0.302</td>
<td>–1.1%</td>
</tr>
<tr>
<td></td>
<td>( \theta_{clock}^* )</td>
<td>50.3%</td>
<td>0.601</td>
<td>59.1%</td>
</tr>
</tbody>
</table>
Borovsky, 2013b]. Exploration of the input solar wind variables to CCA finds that time averaging the clock angle-type variables improves the first canonical correlation coefficient \(r_{(1)}\), whereas time averaging other solar wind input variables negligibly improves or weakens the coefficients \(r_{(1)}\). More of this time averaging will be seen in section 6.

5. Interpretation of Second Canonical Correlations

The second canonical correlation involves a pair of variables, \(S_{(2)}\) and \(E_{(2)}\), that are uncorrelated with \(S_{(1)}\) and with \(E_{(1)}\) and that have the highest correlation coefficient with each other given those uncorrelation restrictions.

When using CCA to compare solar wind data sets to geomagnetic indices data sets, very often a pattern is seen wherein the second-canonical-correlation composite Earth variable forms with a difference between the magnitudes of the high-latitude indices versus the global-convection indices. Specifically, the Earth variable \(E_{(1)}\) of the first canonical correlation is of the form of a sum of the geomagnetic indices such that the coefficients of \(AE\), \(-AL\), \(AU\), PCI, \(Kp\), and \(-MBI\) are all positive, and the Earth variable \(E_{(2)}\) of the second canonical correlation has positive coefficients of \(Kp\) and \(-MBI\) and negative coefficients of \(AE\), \(-AL\), \(AU\), and PCI. Hence, \(E_{(1)}\) is additive in the geomagnetic indices with all indices contributing positively as geomagnetic activity increases, and \(E_{(2)}\) contains a differentiation between the global-convection indices \((Kp\) and \(-MBI\)) and the high-latitude indices \((AE\), \(-AL\), \(AU\), and PCI\). Simultaneously, the pattern of behavior of \(S_{(1)}\) and \(S_{(2)}\) is as follows: the signs of the coefficients of the solar wind variables tends to be the same in \(S_{(1)}\) and \(S_{(2)}\) except for clock angle-type variables such as \(\theta_{clock}\), \(\sin^2(\theta_{clock}/2)\), and \(B_z\) which are reversed between \(S_{(1)}\) and \(S_{(2)}\).

An example of this pattern is given in Table 5. Here the solar wind variables are chosen to be \(\log(nvsw^2)\), \(\log(n)\), \(\log(Ma)\), and \(\theta_{clock}\), and the geomagnetic indices are chosen as the simple set \(\log(AE_1)\), \(PCI_0\), and \(Kp_1\). In Table 5 note the reversal of the sign of \(\theta_{clock}\) and the reversals of the signs of \(\log(AE_1)\) and \(PCI_0\) between the “first” and “second” rows.

### Table 4. The First Canonical Correlation Coefficient \(r_{(1)}\) (Last Column) for Various Combinations of the Solar Wind Inputs as Marked by a Cross

<table>
<thead>
<tr>
<th>Solar Wind Input Variables</th>
<th>Geomagnetic Indices</th>
<th>Canonical Correlation Coefficient (r_{(1)})</th>
<th>(1 - r_{(1)^2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\log(nvsw^2))</td>
<td>(\log(n))</td>
<td>(\log(Ma))</td>
<td>(\sin^2(\theta_{clock}/2))</td>
</tr>
<tr>
<td>1</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>6</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>7</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>8</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>9</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>10</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>11</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>12</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>13</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

\(\text{Note the signs of the coefficients. In the final column the first and second canonical correlation coefficients are listed.}\)
For analysis of the first and second canonical correlations of Table 5, let us define the following four variables from the coefficients of Table 5:

\[
\begin{align*}
\text{HILAT} &= \log(AE_1)^* + 0.450 \text{ PCI}_0^* \\
\text{CONVECT} &= Kp_1^* \\
\text{STRENGTH} &= \log(n_{sw}^2)^* - 0.660 \log(n)^* - 0.308 \log(M_a)^* \\
\text{CLOCK} &= \theta_{\text{clock}}^* 
\end{align*}
\]

(7a) (7b) (7c) (7d)

where HILAT represents the magnitude of the high-latitude geomagnetic indices, CONVECT represents the magnitude of the global-convection indices, STRENGTH represents the driver strength of the solar wind, and CLOCK represents the clock angle of the solar wind magnetic field. The values of HILAT, CONVECT, STRENGTH, and CLOCK can be obtained as functions of time using expressions (7a)–(7d) and the OMNI2 data set (with 102,900-hourly values of each of the four variables generated using expressions (7a)–(7d) and 102,900 h of solar wind data). Using the time series of STRENGTH and CLOCK as two solar wind input variables and using the time series of HILAT and CONVECT as two geomagnetic index input variable, the canonical correlation analysis of these variables yields

\[
\begin{align*}
S_{(1)}(t) &= 0.800 \text{ STRENGTH} + 0.610 \text{ CLOCK} \quad (8a) \\
E_{(1)}(t) &= 0.586 \text{ HILAT} + 0.468 \text{ CONVECT} \quad (8b) \\
S_{(2)}(t) &= 0.601 \text{ STRENGTH} - 0.792 \text{ CLOCK} \quad (8c) \\
E_{(2)}(t) &= -1.556 \text{ HILAT} + 1.596 \text{ CONVECT} \quad (8d)
\end{align*}
\]

with canonical correlation coefficients \(r_{(1)} = 0.841\) and \(r_{(2)} = 0.502\). In a sense, \(S_{(1)}\) is a fit to \(E_{(1)}\) (something to describe \(E_{(1)}\)) and in a sense \(S_{(2)}\) is a fit to \(E_{(2)}\) (something to describe \(E_{(2)}\)). If we set \(E_{(1)} = S_{(1)}\) and set \(E_{(2)} = S_{(2)}\) and use expressions (8a)–(8d), we get the following two equations:

\[
\begin{align*}
0.468 \text{ CONVECT} + 0.586 \text{ HILAT} &= 0.800 \text{ STRENGTH} + 0.610 \text{ CLOCK} \quad (9a) \\
1.596 \text{ CONVECT} - 1.556 \text{ HILAT} &= 0.601 \text{ STRENGTH} - 0.792 \text{ CLOCK} \quad (9b)
\end{align*}
\]

Expression (9a) comes from the first canonical correlation, and expression (9b) comes from the second canonical correlation. The two algebraic expressions (9a) and (9b) can be solved for HILAT and CONVECT, which yields

\[
\begin{align*}
\text{HILAT} &= 0.60 \text{ STRENGTH} + 0.81 \text{ CLOCK} \quad (10a) \\
\text{CONVECT} &= 0.96 \text{ STRENGTH} + 0.29 \text{ CLOCK} \quad (10b)
\end{align*}
\]

Examining expressions (10a) and (10b), it can be seen that the HILAT geomagnetic indices have a stronger dependence on CLOCK than do the CONVECT geomagnetic indices. One might presume that this is owed to the 3 h averaging that goes into \(Kp\) (which makes up CONVECT) which would decrease its dependence on the 1 h averaged value of the rapidly varying CLOCK; however, if the 1 h resolution MBI index is used instead of \(Kp\), this reduced dependence on the clock angle of the global-convection indices is still seen.

In comparing the solar wind data set with the geomagnetic index data set, the first canonical correlation yields information about which solar wind variables are correlated with geomagnetic activity in general and information about the strength of those correlations; the second canonical correlation provides information about the difference in solar wind driving of high-latitude geomagnetic indices versus magnetospheric-convection geomagnetic indices.

### 6. Generating and Examining Solar Wind Driver Functions for the Magnetosphere

By using CCA to compare multivariate solar wind data sets to multivariate geomagnetic index data sets, the solar wind data set can be used to predict geomagnetic activity, in general, (sort of a geomagnetic index stock market prediction) via a canonical solar wind driver function. It will turn out that the canonical solar wind driver functions that CCA generates are also superior for describing the individual geomagnetic indices. Hence, the driver functions generated are in a sense universal.
An accurate and relatively simple driver function can be generated by using the seven geomagnetic indices as inputs along with the solar wind variables \( \log(n_{sw}^2) \), \( \log(n) \), \( \log(M_A) \), and \( \sin^2(\theta_{\text{clock}}/2) \) (cf. section 3 and Table 2). Note that \( \log \) is the natural (base e) logarithm. With these inputs, the CCA procedure yields the first canonical solar wind variable

\[
S_{(1)} = 1.196 \log(n_{sw}^2)^* - 0.784 \log(n)^* - 0.388 \log(M_A)^* + 0.604 \sin^2(\theta_{\text{clock}}/2)^*.
\]

(11)

To make this canonical solar wind variable into a useful solar wind driver function, we convert the dimensionless standardized variables in expression (11) into dimensional unnormalized variables. For instance, \( \log(n)^* = (\log(n) - 1.721)/0.681 \), so the term \( 0.784 \log(n)^* \) becomes \( (0.784/0.681) \log(n) \) plus a constant. Because we are only interested in correlations of variances, the constants can be ignored when unnormalizing the variables. Dividing by the coefficients that multiplies the \( \log(n_{sw}^2)^* \) term, expression (11) yields the CCA-generated driver function

\[
C_1 = \log(n_{sw}^2) - 0.561 \log(n) - 0.445 \log(M_A) + 0.958 \sin^2(\theta_{\text{clock}}/2)
\]

(12)

In expression (12) \( n_{sw}^2 \) is in units of \( \text{cm}^{-3} \cdot \text{km}^2/\text{s}^2 \), \( n \) is in units of \( \text{cm}^{-3} \), and \( M_A \) and \( \sin^2(\theta_{\text{clock}}/2) \) are dimensionless. The first canonical correlation coefficient for this function is \( r_{(1)} = 0.859 \) between \( C_1 \) and \( E_{(1)} \). The Pearson linear correlation coefficients between the driver function \( C_1 \) and the seven individual geomagnetic indices are listed in Table 6. The average of the correlation coefficient between \( C_1 \) and the seven indices is 0.723. As can be seen by comparing \( C_1 \) with the Newell function \( \nu_{sw}^{4/3}B_z^{-2/3}\sin^{8/3}(\theta_{\text{clock}}/2) \), with \( R_{\text{quick}} \), and with \( R_2 \) in Table 6, this simple function \( C_1 \) does a better job of predicting the overall variance of individual geomagnetic indices: not quite as good on the high-latitude indices \( AE \) and \( AL \) but substantially better on the global-convection indices \( Kp \) and \( MBI \).

Repeating the above procedure using the 2 h average of \( \sin^2(\theta_{\text{clock}}/2) \) (the original hour plus the hour beforehand), denoted as \( < \sin^2(\theta_{\text{clock}}/2) >_2 \), yields the solar wind driver function

\[
C_2 = \log(n_{sw}^2) - 0.556 \log(n) - 0.454 \log(M_A) + 1.174 < \sin^2(\theta_{\text{clock}}/2) >_2
\]

(13)

The first canonical correlation coefficient for this function is \( r_{(1)} = 0.879 \) between \( C_2 \) and \( E_{(1)} \). The correlation coefficients between \( C_2 \) and the seven individual geomagnetic indices are listed in Table 6: the seven-index average of the correlation (which is defined as the average of the seven individual correlations for the seven indices) for \( C_2 \) is 0.735. As can be seen in Table 6, the correlations with \( AE \) and \( AL \) are reduced from \( C_1 \) to \( C_2 \) with the introduction of the time averaging of \( \sin^2(\theta_{\text{clock}}/2) \), and the correlations with the other five indices go up. This could be interpreted as either (1) the indices \( AE \) and \( AL \) are more sensitive to the instantaneous value of \( \theta_{\text{clock}} \) than the other indices are or (2) the CCA’s emphasis of describing variance between the solar wind input variables and the geomagnetic index input variables has shifted to different geomagnetic indices. In support of interpretation (1), if direct linear correlations are calculated between the seven indices and \( \sin^2(\theta_{\text{clock}}/2) \), time-averaging \( \sin^2(\theta_{\text{clock}}/2) \) improves five of the correlations but reduces the correlations with \( \pi AL \) and \( AE \). Also in support of interpretation (1), the analysis of the first and second canonical correlations in section 5 found that the high-latitude indices respond to information about the clock angle much more than the global-convection indices do (cf. expressions (10a) and (10b)). Contradicting interpretation (1), one would think that the polar cap

<table>
<thead>
<tr>
<th>Geomagnetic Indices</th>
<th>( AE )</th>
<th>( -Al )</th>
<th>( AU )</th>
<th>( PCI )</th>
<th>( Kp )</th>
<th>( -MBI )</th>
<th>( -Dst^2 )</th>
<th>Seven-index average</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_{\text{quick}} )</td>
<td>76.6</td>
<td>74.4</td>
<td>70.1</td>
<td>69.4</td>
<td>70.1</td>
<td>67.7</td>
<td>65.6</td>
<td>70.6</td>
</tr>
<tr>
<td>( R_2 )</td>
<td>74.4</td>
<td>71.0</td>
<td>66.9</td>
<td>69.4</td>
<td>70.5</td>
<td>68.2</td>
<td>65.5</td>
<td>72.3</td>
</tr>
<tr>
<td>( C_1 )</td>
<td>74.4</td>
<td>71.0</td>
<td>66.9</td>
<td>68.2</td>
<td>70.5</td>
<td>65.5</td>
<td>72.3</td>
<td>73.5</td>
</tr>
<tr>
<td>( C_2 )</td>
<td>74.1</td>
<td>71.0</td>
<td>66.9</td>
<td>68.2</td>
<td>70.5</td>
<td>65.5</td>
<td>72.3</td>
<td>78.2</td>
</tr>
<tr>
<td>( C_3 )</td>
<td>79.9</td>
<td>75.2</td>
<td>72.8</td>
<td>79.3</td>
<td>80.9</td>
<td>86.4</td>
<td>70.8</td>
<td>78.2</td>
</tr>
</tbody>
</table>

The bottom row is the seven-index average of the coefficients (the average of the individual correlations for the seven different indices) for each driver function.
index PCI with its shortest time lag would be the most sensitive to the current value of $\theta_{\text{clock}}$, but its correlation increased substantially with the introduction of time averaging of $\sin^2(\theta_{\text{clock}}/2)$.

As noted in Table 4, adding more solar wind variables into the CCA procedure increases the first canonical correlation coefficient and leads to a more accurate solar wind driver function. Applying the CCA procedure to the seven geomagnetic index inputs plus 11 solar wind inputs of the bottom row of Table 4, and using a 3 h averaged $<\sin^2(\theta_{\text{clock}}/2)>_3$ function (the value at the original hour plus the values in the two previous hours) yields the driver function

$$C_3 = \log(nv_{sw}) - 0.424 \log(n) - 0.451 \log(M_h) + 1.285 <\sin^2(\theta_{\text{clock}}/2)>_3$$
$$- 0.107 B_z + 0.165 \log(F_{10.7}) + 3.01 \times 10^{-3} \theta_{\text{bn}} + 0.241 B/B + 0.0445 B_{\text{mag}} + 2.21 \times 10^{-3} v_{sw}$$

(14)

The solar wind inputs going into expression (14) were chosen by trial and error; input variables that had no effect on the first canonical correlation coefficient were eliminated, and the 11 input variables of expression (14) were the remaining. In expression (14) $B_z$ and $B_{\text{mag}}$ are in units of nT, $F_{10.7}$ is in units of solar flux, $\theta_{\text{bn}}$ is in degrees, $v_{sw}$ is in units of km/s, and $B/B$ and $\alpha/\pi$ are dimensionless. The first canonical correlation coefficient for this function is $r(1) = 0.917$ between $C_3$ and $E_{1}$, with an unaccounted for variance $1 - r(1)^2$ of 15.9%.

(Note that $C_3$ differs from the last row of Table 4 in that $C_3$ utilizes $<\sin^2(\theta_{\text{clock}}/2)>_3$, whereas the last row of Table 4 utilizes $\sin^2(\theta_{\text{clock}}/2)$. As seen in Table 6, the seven-index average of the correlation coefficients (the average of the seven individual correlations) for $C_3$ is 0.782, a substantial improvement over the other driver functions in Table 6. As can be seen in Table 6, the driver function $C_3$ strongly emphasizes the global-convection indices but does good jobs of describing the variance of all of the indices. Note in Table 6 that the correlation coefficient between $C_3$ and $AE_1$ is 0.799; the correlation coefficient between $C_3$ and $\log(AE1)$ (or equivalently between $\exp(C_3)$ and $AE_1$) is 0.819.

Figure 4. A comparison of the (left) Newell driver function, (middle) the reconnection driver function $R_2$, and (right) the $C_3$ driver function describing (top row) the midnight boundary index $MBI_1$ and (bottom row) the $AE_1$ index. In all panels, 102,900-hourly averages are plotted.
In Figure 4 the midnight boundary index MBI1 (top row) is plotted as a function of the Newell driver function (left), the reconnection driver function R2 (middle), and as a function of the driver function C3 (expression (14)) (right). The Pearson linear correlation coefficients are indicated in red on the three panels (see also Table 6). Note in Figure 4 (right) the superior tracking of MBI versus C3, whereas there are systematic differences in the left and middle columns, particularly for low levels of driving. In Figure 4 (bottom row) the auroral electrojet index AE1 is plotted as a function of Newell (left), R2 (middle), and C3 (right). Note again the superior tracking by C3. In particular, for low levels of activity the C3 function provides a superior description of geomagnetic activity.

By using solar wind driver functions as inputs into the CCA procedure, the properties of the driver functions can be examined. In particular, driver functions can be compared with each other via the second canonical correlation, which tends to produce composite canonical variables that involve differences between correlated variables. In Table 7 the coefficients of the first and second canonical variables are listed for the solar wind driver functions R2 [Borovsky, 2013a] and Newell (=v_{sw}^{4/3}B_{⊥}^{2/3}\sin^{2/3}(\theta_{\text{clock}}/2)) [Newell et al., 2007] used as solar wind inputs along with the seven geomagnetic indices used as Earth inputs. The second row of Table 7 is the second canonical correlation. As can be seen in the second row S_{(2)} = 1.00, Newell − 1.01 R_{2}^* is a difference between the two standardized driver functions. The CCA process aims to maximize the correlation between S_{(2)} and E_{(2)}; large coefficients in E_{(2)} indicate that is where the correlations with S_{(2)} will be greatest. The second row of Table 7 indicates that the Newell driver function exceeds the R2 driver function in accounting for −AL (which has a strong positive coefficient in the second row) and the second row indicates that the R2 driver exceeds the Newell driver in accounting for KP (which has a strong negative coefficient in the second row). Hence, the Newell (electric field based) driver function emphasizes high-latitude indices, and the R2 (reconnection based) driver function emphasizes magnetospheric global convection. It has been argued [Goertz et al., 1993] that the intensity of auroral electrojet indices can be calculated from a mapping of the solar wind electric field along open magnetic field lines into the magnetosphere with a time lag.

### 7. Summary

In this report canonical correlation analysis has been applied to the multivariate solar wind data set and the multivariate geomagnetic index data set to investigate the mathematics and physics of the solar wind driving of the magnetosphere-ionosphere system. The findings of this study are numbered as follows:

1. The concept of an Earth vector of geomagnetic indices was introduced rather than studying the reaction of one geomagnetic index at a time. This Earth vector or Earth variable provides a more system-wide measure of activity in the magnetosphere. An analogy is the stock market index rather than an individual stock.
2. Examining the first canonical correlations between the solar wind data set and the geomagnetic index data set yielded information about which quantities in the solar wind contribute to the coupling. A question was raised (but not settled) as to which is superior: the use of primitive solar wind variables as inputs or the use of physical solar wind quantities as inputs.
3. Examining the second canonical correlations in comparison with the first canonical correlations yielded information about the difference in solar wind driving of high-latitude geomagnetic indices (e.g., AE and PCI) versus magnetospheric global-convection geomagnetic indices (e.g., KP and MBI). The high-latitude indices react more strongly to the clock angle of the solar wind magnetic field than do the global-convection indices, and the global-convection indices react more strongly to the strength of the solar wind driving function than do the high-latitude indices.
4. Canonical correlation analysis does a better job at describing global-convection geomagnetic indices than it does at describing high-latitude geomagnetic indices. This may indicate that there is more predictable variance in the convection indices than there is in the high-latitude indices.

5. Time averaging the clock angle input of the solar wind yields improved canonical correlations between the solar wind data set and the geomagnetic index data set and yields improved descriptions of the variances of most geomagnetic indices. Two exceptions are $AE$ and $AL$ where clock angle time averaging reduces the prediction of the variance of those two indices.

6. Solar wind driver functions for the magnetosphere were generated using the CCA technique. Two simple (four input) driver functions and one more complicated (11-input) driver functions were presented. All three driver functions are superior to existing electric field driver functions and to existing reconnection driver functions. In particular, the functional form of the CCA driver function $C_3$ is superior at low levels of driving and low levels of geomagnetic activity.

7. Canonical correlation analysis was used to examine the differences between a high-quality electric field driver functions and a high-quality reconnection driver functions. The CCA indicated that the electric field-based function emphasized high-latitude indices, and the reconnection-based function emphasized global-convection indices.

8. The canonical correlation analysis provides information about just how much variance of the geomagnetic index is predictable. Without great exploration of inputs and averaging, without adjusting the time lags, and without detrending diurnal and annual signals in the geomagnetic indices that are not in the solar wind, a correlation coefficient of $r = 91.7\%$ between a solar wind driver function and its canonical Earth vector has been obtained using 1 h resolution solar wind measurements and 1 h resolution geomagnetic indices. For this case the unpredicted variance $1 - r^2$ is 15.9%. Using some of the methods described in section 8, this number should be improved upon.

8. **Future Work**

Canonical correlation analysis has opened up a new avenue to study the mathematics and physics of solar wind/magnetosphere coupling. This report only contains the rudimentary application of CCA and early findings. Below is a brief list of possible expansions of the use of CCA for the study of the solar wind-driven magnetosphere-ionosphere system.

More measures of the state of the magnetosphere-ionosphere system are needed to fully exploit the capabilities of the CCA technique. In a subsequent report, magnetospheric ULF indices are incorporated [Borovsky and Denton, 2014]. In future a measure of the open flux in the polar cap could be made available and added into the CCA process. Missing and needed are long-term continuous measures of (a) the plasma density in the dayside and the nightside magnetosphere, (b) the rate of mass flow through the magnetosphere, (c) the rate of ionospheric ion outflow, (d) the oxygen fraction of the magnetospheric plasmas, (e) the state of the outer electron radiation belt, and (e) the properties of the substorm-injected particle populations.

A systematic examination of the use of primitive solar wind variables versus physical solar wind variables needs to be performed to try to discern which are more fundamental to describing the mathematics and physics of solar wind/magnetosphere coupling.

In the present study fixed time lags were used between the time of solar wind measurements (propagated to the front of the magnetosphere) and the various geomagnetic indices. Those fixed lags (with 1 h time resolution) were determined in a prior study using the first-generation reconnection driver function $R$ [Borovsky, 2008]. Using variable time lags (which depend, for instance, on the level of geomagnetic activity) should improve the performance of the canonical correlation analysis between the solar wind and the magnetosphere.

Examining solar wind and geomagnetic index autocorrelation functions, Borovsky [2013a] uncovered diurnal and annual trends in the geomagnetic indices that are not in the solar wind. Using the residuals between solar wind driver functions and the geomagnetic indices, these diurnal and annual trends can be fit and removed from the geomagnetic indices. This yields an improved description of the variance of the geomagnetic indices by the variances in the solar wind. Applying these detrending techniques in coordination with CCA should prove improved driver functions and improved descriptions of the Earth vector by the solar wind.
Hints about differences in the way the solar wind drives magnetospheric convection versus how it drives high-latitude (auroral) activity were found in the present CCA study. A more thorough investigation of what CCA can yield about the differences is needed.

The first and second canonical correlations between the solar wind data set and the geomagnetic data set were briefly explored. The second canonical correlation yielded information about the different reactions to the solar wind of high-latitude activity versus magnetospheric global convection. Examining the third canonical correlations may yield information about differences between the driving of the AL and AU indices (dayside versus nightside current systems), or it may yield information about the differences of the driving of the Dst index (equatorial plasma pressure and cross-tail current) and the high-latitude and global-convection indices.

The driver functions generated by the CCA technique in this report had solar wind input variables in the forms of logarithms. These logarithmic terms make physical interpretation of the solar wind driving nonobvious. With more work, driver functions with more appealing functional forms can be generated.

Plots of Earth variables as functions of solar wind variables make it obvious that the coupling between the solar wind and the magnetosphere is nonlinear. Exploring functional forms of the solar wind input variables can uncover and account for these nonlinearities.

Finally, the CCA technique can be coupled to other techniques (such as the mutation algorithm mentioned in section 1 that generates solar wind driver functions) to produce yet-more-powerful mathematical function to describe the solar wind driving of the magnetosphere-ionosphere system and to uncover the physical processes underlying that coupling.

Acknowledgments

The authors wish to thank Derek Bingham, Joachim Birn, Mick Denton, and Bob McPherron for useful conversations and Jung-Chao Wang for use of a matrix eigenvalue algorithm. This work was supported at Space Science Institute by the NASA CCMSS-24 Program, the NSF GEM Program, and the NASA Heliophysics Guest Investigator Program at the University of Michigan by the NASA Geospace SBST Program and at the University of Lancaster by Science and Technology Funding Council grant ST/000801/1.

Yuming Wang thanks Nikolai Ostgaard and another reviewer for their assistance in evaluating the paper.

References


ONimon, K., R. K. Henson, and M. S. Gates (2010), Revisiting interpretation of canonical correlation analysis: A tutorial and demonstration of

Thomsen, M. F., J. E. Borovsky, D. J. McComas, and M. R. Collier (1998), Variability of the ring current source population,


Space Sci., 17, 1095–1102.


