Gap filling of solar wind data by singular spectrum analysis

D. Kondrashov,[1] Y. Shprits,[1] and M. Ghil[1,2]

Received 28 May 2010; accepted 17 June 2010; published 4 August 2010.

[1] Observational data sets in space physics often contain instrumental and sampling errors, as well as large gaps. This is both an obstacle and an incentive for research, since continuous data sets are typically needed for model formulation and validation. For example, the latest global empirical models of Earth's magnetic field are crucial for many space weather applications, and require time-continuous solar wind and interplanetary magnetic field (IMF) data; both of these data sets have large gaps before 1994. Singular spectrum analysis (SSA) reconstructs missing data by using an iteratively inferred, smooth "signal" that captures coherent modes, while "noise" is discarded. In this study, we apply SSA to fill in large gaps in solar wind and IMF data, by combining it with geomagnetic indices that are time-continuous, and generalizing it to multivariate geophysical data consisting of gappy "driver" and continuous "response" records. The reconstruction error estimates provide information on the physics of co-variability between particular solar-wind parameters and geomagnetic indices. **Citation:** Kondrashov, D., Y. Shprits, and M. Ghil (2010), Gap filling of solar wind data by singular spectrum analysis, *Geophys. Res. Lett.*, *37*, L15101, doi:10.1029/2010GL044138.

1. Introduction

[2] The main historical—i.e., pre-1994—solar-wind and interplanetary magnetic field (IMF) observations come from measurements taken on board of the IMP-8 spacecraft. While the spacecraft crossed the magnetosheath and magnetosphere, it was not immersed in the solar wind, and so large continuous gaps exist in the collected data.

[3] *Qin et al.* [2007] developed a decorrelation-time-based approach to interpolate the solar-wind characteristics across data gaps, and to evaluate parameters needed for global empirical magnetic models, like that of *Tsyganenko and Sitnov* [2005]. For very large gaps, however—such as those in 1990–1991, whose lengths far exceed typical decorrelation times—this method only yields the mean values of solar-wind parameters, with scantly any variability.

[4] The behavior of Earth's magnetosphere is strongly influenced by the solar wind. Various geomagnetic indices—such as *Kp*, D_{st} or *AE*—are inferred from ground-measured, and hence time-lagged magnetic disturbances that are caused by the magnetosphere's interaction with the solar wind and the embedded IMF; these indices are typically available continuously in time, even when solar-wind data are not. Broadly speaking, these indices can be considered as a proxy for the overall time-lagged *magnetospheric response* to the *solar driver*, i.e., to the solar wind and IMF.

[5] The purpose of this paper is to reconstruct data in the gaps of the solar driver by using smooth spatio-temporal modes of co-variability inferred by singular spectrum analysis (SSA) from time-lagged correlations in multivariate data—the data sets consisting of various geomagnetic indices, solar wind and IMF—while discarding the noise. *Kondrashov and Ghil* [2006] developed an SSA-based gap-filling method with applications that rely, so far, mainly on the presence of significant oscillatory modes in the time series [see *Kondrashov et al.*, 2005]. We show here that this method can also be successfully applied to multivariate geophysical data sets that consist, broadly speaking, of gappy-driver and continuous-response records, by relying on analogous episodes of co-variability which is not necessarily periodic.

[6] Regression based techniques [*Vassiliadis et al.*, 1995; *Chen and Sharma*, 2006] can be used to derive predictive filter that relate time series of lagged input from the past, i.e., solar wind parameters, to the current values of magnetospheric output, such as geomagnetic indices. This work differs in two key aspects: (i) by inferring the gappy driver from a continuous response, (ii) strictly speaking, it is not a prediction; SSA finite-impulse response filter (FIR, see section 2) is two-sided symmetric and thus uses information from both the "past" and the "future" of dominant modes of driver-response co-variability, derived by extending principal component analysis to the time domain.

[7] In section 2, we briefly review the SSA formulation for a continuous time series and the gap-filling methodology. In section 3, we show very promising results in applying SSA to fill synthetic gaps in hourly solar-wind and IMF data. The paper concludes with a summary of the results in section 4.

2. Data and Methods

2.1. Data

[8] In this study we used publicly available solar-wind and IMF data from the OMNIWEB database at http://omniweb.gsfc.nasa.gov/. This data set combines observations from multiple spacecraft and is appropriately time-shifted to take into account their spatial separation with respect to Earth; *Kp* and D_{st} indices were obtained from the World Data Center (WDC) for Geomagnetism, WDC-Kyoto. The planetary 3-hour *Kp* index is a magnetic-activity measure averaged from several geomagnetic observatories worldwide; it is given on a quasi-logarithmic scale from 0 to 9. The disturbance storm time index D_{st} is used to assess the severity of magnetic storms; it is based on the average value of the horizontal component of the Earth's magnetic field measured hourly at four near-equatorial geomagnetic observatories. Even though *Kp* is a 3-hour index, it was

[1]Department of Atmospheric and Oceanic Sciences and Institute of Geophysics and Planetary Physics, University of California, Los Angeles, California, USA.
[2]Geosciences Department and Laboratoire de Météorologie Dynamique, Ecole Normale Supérieure, CNRS, IPSL, Paris, France.

0094-8276/10/2010GL044138

given an hourly resolution to match the sampling of solar-wind, IMF and D_{st}.

2.2. Gap Filling Methodology

[9] Classical SSA [*Vautard and Ghil*, 1989; *Ghil et al.*, 2002] is a data-adaptive, nonparametric method for spectral estimation based on embedding multivariate time series $\{X(t,l)$: $t = 1,\ldots,N$; $l = 1,\ldots, L\}$ in a vector space of dimension $M < N$, and for a given window width M, the orthonormal set $\{\mathbf{E}_k$: $k = 1,\ldots, LM\}$ of eigenvectors of $\mathbf{C}_X$— called empirical orthogonal functions (EOFs) — is the *optimal data-adaptive* set that spans the given time series, i.e., for any $1 \le K \le LM$, the set of K leading EOFs captures the maximum variance. Projecting $X(t,l)$ onto each EOF yields the corresponding principal component (PC) $\mathbf{A}_k$; the entire time series or parts thereof can be reconstructed by using linear combinations of PCs and EOFs, which yield the reconstructed components (RCs) $\mathbf{R}_k$.

[10] The SSA gap-filling procedure of *Kondrashov and Ghil* [2006] uses temporal correlations in the data—or spatio-temporal ones in the multivariate case, as used in this study,—to reconstruct the missing points with coherent signal modes, while discarding the noise; it also estimates the power spectrum of a gappy time series.

[11] The procedure consists of two main steps: (i) obtain iteratively estimates of missing values $\hat{X}(t)$ by using the leading subset of RCs, which are then applied to update a self-consistent lag-covariance matrix $\mathbf{C}_X$, EOFs $\mathbf{E}_k$ and PCs $\mathbf{A}_k$; and (ii) use cross-validation to optimize the window width M^* and number of "signal" modes K^* to fill the gaps (see below).

[12] First, the original data set is centered by computing the unbiased value of the mean and setting the missing-data values to zero. The inner-loop iteration starts by computing the leading EOF $\mathbf{E}_1$ of the centered, zero-padded record. The corresponding RC $\mathbf{R}_1$ is used next to obtain nonzero values in place of the missing points; the new record's mean, covariance matrix and EOFs are then recomputed by SSA. The reconstruction of the missing data is repeated with a new estimate of $\mathbf{R}_1$ until a convergence test has been satisfied; in the present application we use 2.5% of normalized root-mean-square (rms) error as a criterion.

[13] The objective of the outer-loop iteration is to separate the signal from the noise. To start it, we add $\mathbf{E}_2$ to the reconstruction, by using the solution with data filled in by $\mathbf{R}_1$, and repeat the inner iteration with two EOFs until it converges; then another EOF is added and so on. We stop the outer iteration once K^* modes attributed to signal, have been processed, and higher ranked modes are assumed to be noise.

[14] A useful way to look at SSA gap-filling is in terms of applying iteratively data-adaptive finite-impulse response filters (FIR); each reconstruction filter $\mathbf{f} = (f_{-M+1}, f_{-M}, \ldots f_{-1}, f_0, f_1, \ldots f_{M-1})$ is symmetric, has a length of $2M - 1$, and represents the combined influence of the EOFs used so far in the outer-loop iteration [*Kondrashov and Ghil*, 2006; *Varadi et al.*, 1999]. For the multivariate data in this study, the gaps of the driver are filled-in mainly by the filtered time series of the continuous response channel, representing smooth modes of co-variability captured by the multidimensional EOFs.

[15] The optimal SSA parameters for gap filling are found by cross-validation experiments: for each experiment, a fixed but randomly chosen fraction of *available* (i.e., excluding *missing*) data is left out, and the rms error in reconstruction is computed as a function of the number K of EOFs retained and of the SSA window size M. The global minimum in error, averaged over all experiments, corresponds to the required optima K^* and M^*, and provides an estimate of the actual error in the reconstructed data set $\hat{X}(t)$.

3. Results and Interpretation

3.1. Choice of Time Interval

[16] As a proof-of-concept, we copy data gaps from 441 days in 1990–1991 to create synthetic gaps in the time series of hourly B_z and P during 441 days in 2000–2001; in the new data set there is a total of 10 584 data points, but roughly 58% of the data set is missing. We apply our SSA gap-filling methodology to this gappy data set, combined with the hourly sampled Kp, the hourly D_{st} index or both; the latter are available continuously during 2000–2001 and we use them to help reconstruct B_z and P in the artificially created gaps. The results of the reconstructed data for synthetic data gaps will be compared to actual values.

3.2. Gap-Filling Results

[17] During the SSA gap-filling iterations, the time series are normalized by their standard deviations to bring different types of data within the same range of values, and the RCs are then renormalized. Since SSA gap-filling is a purely statistical method, a threshold was imposed on minimum values during iterations, to avoid physically unrealistic negative values for the dynamic pressure P.

[18] The reconstructed data set $\hat{X}(t)$ is then compared with known values $X(t)$ in the gaps only (excluding available data) by using correlations *Corr* and normalized rms differences *RMS*. Perfect reconstruction corresponds to $Corr = 1$ and $RMS = 0$, respectively. Since SSA gap-filling, however, discards the noise modes, the skill is less then perfect, at best.

[19] These metrics are shown in Figures 1a–1d as a function of the number of SSA modes included in the reconstruction. The best results—i.e., highest correlation and smallest rms—are obtained with the same number of SSA modes, $K^* = 25$, but different window widths: $M^* = 15$ hr for the IMF component B_z and $M^* = 20$ hr for the dynamic pressure P; other window sizes gave inferior results (not shown). The skill metrics are quite good for both B_z and P when using optimal parameter values K^* and M^*, with an edge for dynamic pressure, with $Corr = 0.87$ and $RMS = 0.47$. Making use of the two indices Kp and D_{st} together yields a noticeable improvement for the reconstruction of P, while it makes but little difference in B_z, since most of the skill for the latter comes from the D_{st} index.

[20] Optimal SSA parameter values M^* and K^* are obtained empirically; since SSA is based on optimal decomposition of variance within a sliding time window, the resulting M^* of 15 and 20 hours roughly correspond to the timsescale of the main phase of a storm when geomagnetic indices vary the most. The K^* leading modes capture "useful variance" for gap-filling (see section 2.2) from maximum of $M \cdot L$ modes, where L is number of channels in dataset; in this study $L = 3$ when both Kp and D_{st} are used to fill-in either B_z or P; $L = 1$ for univariate SSA case. Relatively large value of K^* indicates that there is no strong separation

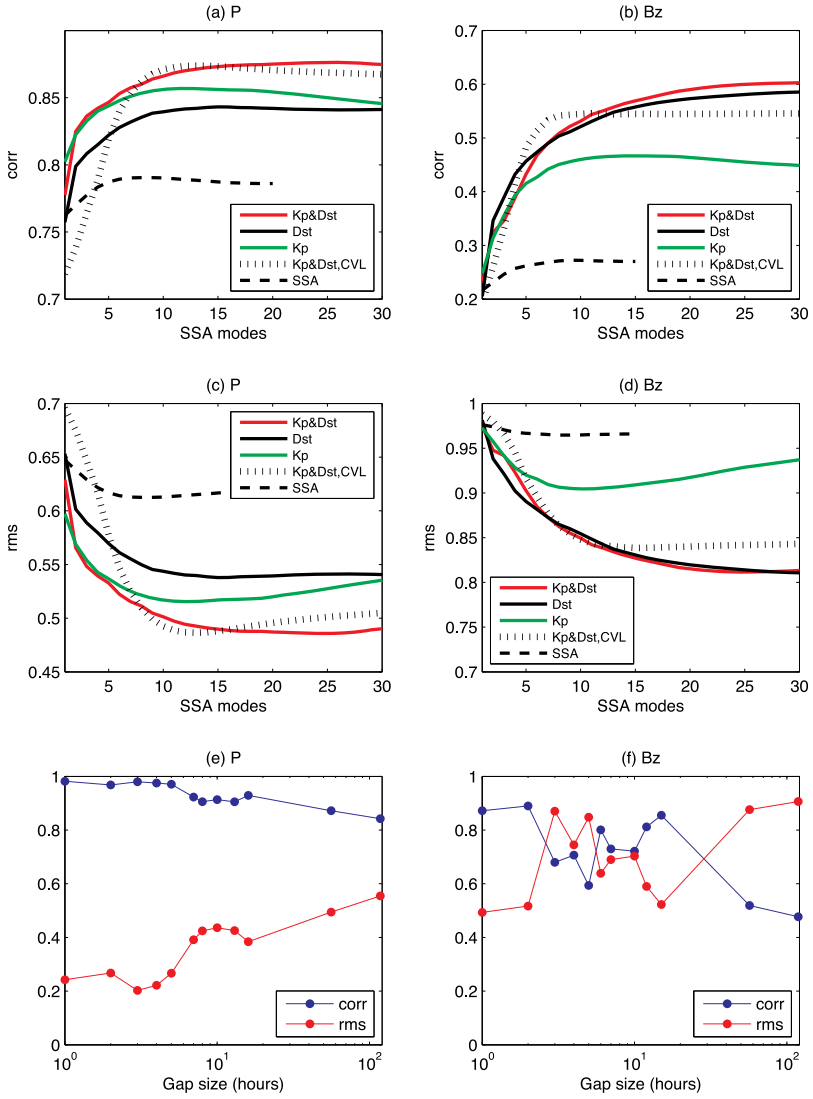

Figure 1. (a–d) Reconstruction skill (correlations *Corr* and rms errors *RMS*) with $M^* = 15$ hr for B_z and $M^* = 20$ hr for P, computed over all gaps as the number of M-SSA modes increases: using Kp only (green), D_{st} only (black), both indices (red), univariate data without indices (dashed black), and estimated skill from cross-validation (dotted black). (e and f) Skill computed for data in gaps of the same size, using both Kp and D_{st} with $K^* = 25$ modes.

between coherent modes of variability and noise; detailed analysis of leading SSA modes is beyond the scope of this paper.

[21] When univariate SSA gap-filling is applied to P and B_z using the same M^*, reconstruction skill is much worse, see dashed black line in Figures 1a–1d (similar result is obtained for other window sizes, not shown). In this case successful reconstruction relies on presence of statistically significant coherent modes inferred by SSA [*Kondrashov and Ghil*, 2006], which are less prominent in B_z then in P. These results confirm that SSA gap-filling can be substantially improved by combining gappy-driver and continuous-response records, and relying on analogous episodes of co-variability which are not necessarily periodic, see more discussion in section 4.

[22] Our results suggest that optimal reconstruction for other solar-wind and IMF data may thus be achieved with different geomagnetic indices and SSA windows, while it may be necessary to systematically search for optimal SSA parameters and combinations of geomagnetic indices for other temporal resolutions.

[23] The time series of filled-in B_z and P (blue line) and original data (red line) for the optimal SSA parameters are shown in Figure 2. When B_z is southward, it is successfully reconstructed, which is consistent with it's predominant control of geomagnetic activity and D_{st} in particular [*Gonzalez and Echer*, 2005; *O'Brien and McPherron*, 2000]; the dynamic pressure, on the other hand, is reconstructed consistently well over the entire testing interval.

[24] Dependence of reconstruction skill to the gap size is demonstrated in Figures 1e and 1f, where it is computed for data combined in gaps of similar length. For very large gaps (>1 day), there are typically very few independent gap samples (one or two) available for averaging. Such gaps have been combined to ensure that there are at least 15 independent samples to compute the skill for effective average gap size.

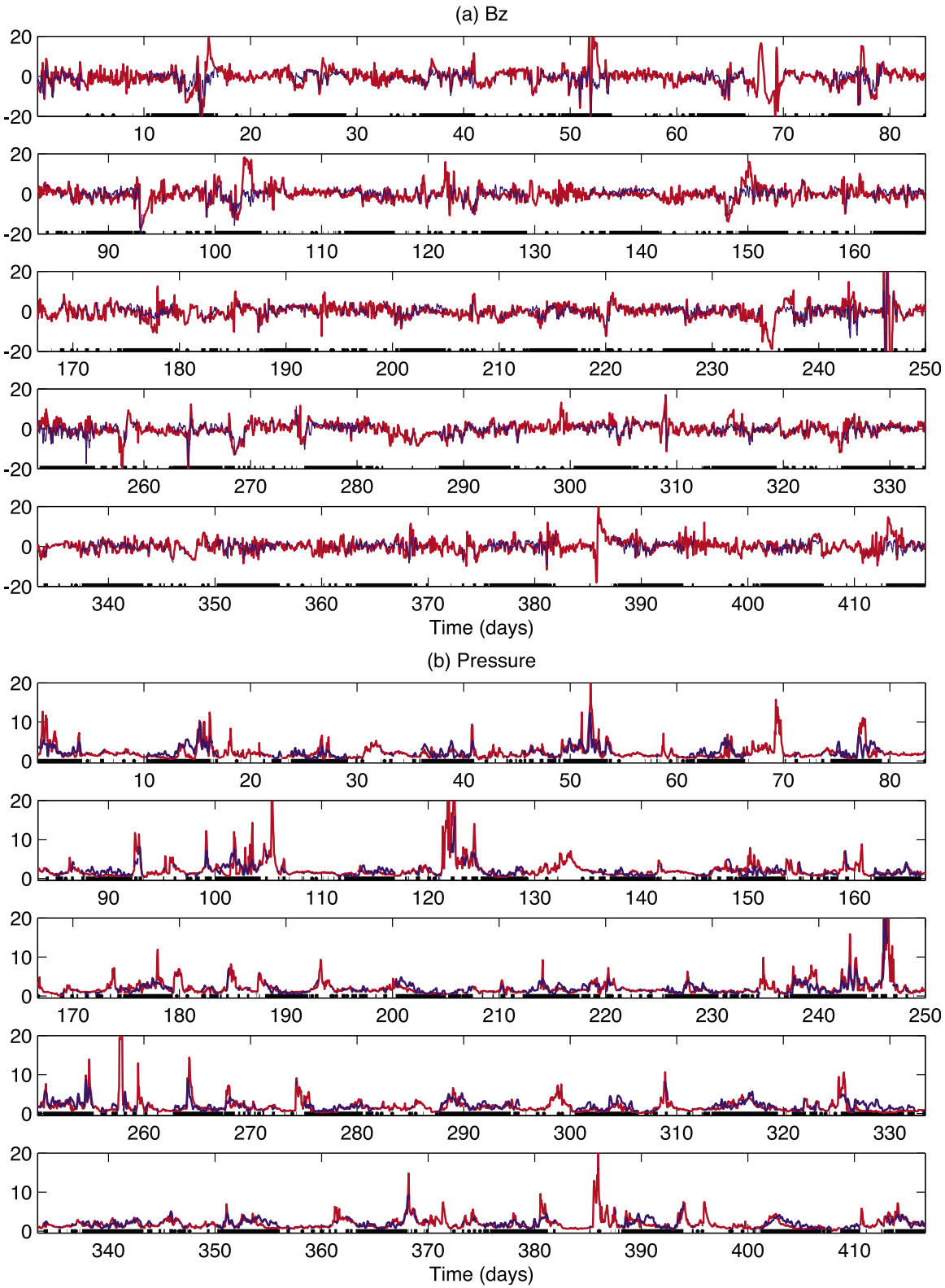

Figure 2. Optimal M-SSA reconstruction (blue curve) in the synthetic gaps (heavy black segments on the time axis) removed from the continuous, hourly 2000–2001 solar-wind data set: (a) B_z, and (b) P; original data in red, see text for details.

As expected, SSA gap-filling works best for smaller gap sizes (<10 hours). The dynamic pressure shows more consistent skill then B_z over all gap sizes, with the relatively high skill scores attained for gaps as large as 5 days.

[25] When applied to historical gappy data, the optimal parameters for gap-filling have to be found through cross-validation (see section 2.2). For our synthetic gaps, these parameters were inferred correctly, and the estimated skill by

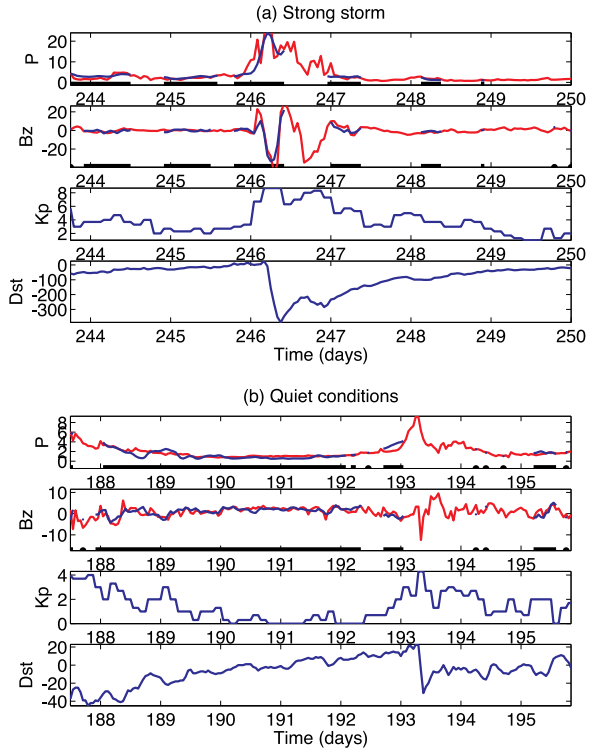

Figure 3. M-SSA gap-filling results for two types of geomagnetic conditions: (a) strong storm, and (b) quiet episode; same conventions as in Figure 2. Also shown are *Kp* and D_{st} variations (blue curves only).

cross-validation is fairly close to the actual values (compare dashed black and red in Figures 1a–1d). In order to obtain such a smooth estimate of the cross-validation curve and accurate estimates of the skill, we selected randomly 80% of the available data (i.e., *excluding synthetic gaps*) in solar driver for each experiment, and averaged results over 10 such experiments. Such a large fraction of data via random selection was needed to obtain large contiguous gaps similar to Figure 2, and thus realistic estimates of the reconstruction skill.

[26] SSA gap-filling works well for a wide range of conditions, as shown in Figure 3 for two selected events during the test period: a very strong geomagnetic storm ($Kp \approx 9$, and $D_{st} < -300$ during days 246–247 (Figure 3a)), and quiet-time magnetosphere (near zero *Kp* and D_{st} during days 188–192 (Figure 3b)). The impressively high quality of the reconstruction inside these selected gaps is due to the ability of SSA to infer smooth temporal modes of co-variability between the solar driver and the response from the existing data. In particular, it is the presence of similar episodes of geomagnetic activity with solar driver data available, that allows SSA to capture such modes via a windowed data-adaptive filter within a leading subset of EOFs of co-variability.

4. Concluding Remarks

[27] We showed here that SSA can be used to fill in large gaps in past solar-wind and IMF data. The novel feature with respect to previous gap-filling applications of SSA is that we considered a gappy driver—the solar wind, represented by the IMF component B_z and dynamic pressure *P*—and a continuously available response; in the present case, the latter was given by the geomagnetic indices *Kp* and D_{st}.

[28] Gaps in the solar-wind and IMF data were filled in by the coherent temporal modes of the solar driver, combined with the geomagnetic response. As a result, we obtained realistic variability in large gaps (see Figure 2)—a considerable improvement over currently available interpolation procedures [e.g., *Qin et al.*, 2007]. The method was shown to work well for both strong geomagnetic storms and quiet conditions (see Figures 3a and 3b). Our estimates of reconstruction error (see Figure 1) provide insight into the physics of covariability between particular solar-wind parameters and geomagnetic indices. In particular, *Kp*, which is a general indicator of geomagnetic activity, is much less successful than D_{st}, which mostly shows the ring current intensity, in reconstructing the IMF's B_z. Both indices however contribute significantly to the successful reconstruction of dynamic pressure *P*.

[29] In previous applications to Nile River floods or sea surface temperatures [e.g., *Kondrashov and Ghil*, 2006], successful reconstruction was largely due to the presence of significant oscillatory modes in the univariate or multivariate time series in which the gaps were being filled in. While there is a well-known ≈27-day recurrence in geomagnetic activity due to so-called corrotating interaction regions (CIRs), such periodicity is most prominent during the declining phase of the solar cycle or near its minimum [*Tsurutani et al.*, 2006]. In order for SSA to realistically reconstruct the solar driver in 2000–2001, near the maximum of the solar cycle, one has to rely mainly on analogous episodes of geomagnetic variability during time intervals when solar wind data are available. It is the covariation in driver and response at times when both are present that allows us to reconstruct the former when only the latter is recorded in the data set. We thus expect that SSA gap filling can be applied to other heliospheric, ionospheric and magnetospheric data sets where there is a gappy record of the driver but a continuous record of the response.

[30] **Acknowledgments.** The basic SSA gap-filling algorithm is available in the SSA-MTM Toolkit (http://www.atmos.ucla.edu/tcd/ssa/). This work is supported by the Lab Research Fee grant, 09-LR-04-116720-SHPY.

References

Chen, J., and A. S. Sharma (2006), Modeling and prediction of the magnetospheric dynamics during intense geospace storms, *J. Geophys. Res.*, *111*, A04209, doi:10.1029/2005JA011359.

Ghil, M., et al. (2002), Advanced spectral methods for climatic time series, *Rev. Geophys.*, *40*(1), 1003, doi:10.1029/2000RG000092.

Gonzalez, W. D., and E. Echer (2005), A study on the peak D_{st} and peak negative B_z relationship during intense geomagnetic storms, *Geophys. Res. Lett.*, *32*, L18103, doi:10.1029/2005GL023486.

Kondrashov, D., and M. Ghil (2006), Spatio-temporal filling of missing points in geophysical data sets, *Nonlinear Processes Geophys.*, *13*, 151–159.

Kondrashov, D., Y. Feliks, and M. Ghil (2005), Oscillatory modes of extended Nile River records (A.D. 622–1922), *Geophys. Res. Lett.*, *32*, L10702, doi:10.1029/2004GL022156.

O'Brien, T. P., and R. L. McPherron (2000), Forecasting the ring current index D_{st} in real time, *J. Atmos. Sol. Terr. Phys.*, *62*, 1295–1299.

Qin, Z., R. E. Denton, N. A. Tsyganenko, and S. Wolf (2007), Solar wind parameters for magnetospheric magnetic field modeling, *Space Weather*, *5*, S11003, doi:10.1029/2006SW000296

Tsurutani, B. T., et al.. (2006), Corotating solar wind streams and recurrent geomagnetic activity: A review, *J. Geophys. Res.*, *111*, A07S01, doi:10.1029/2005JA011273.

Tsyganenko, N. A., and M. I. Sitnov (2005), Modeling the dynamics of the inner magnetosphere during strong geomagnetic storms, *J. Geophys. Res.*, *110*, A03208, doi:10.1029/2004JA010798.

Varadi, F., et al. (1999), Searching for signal in noise by random-lag singular spectrum analysis, *Astrophys. J.*, *526*, 1052–1061.

Vassiliadis, D., A. J. Klimas, D. N. Baker, and D. A. Roberts (1995), A description of the solar wind-magnetosphere coupling based on nonlinear filters, *J. Geophys. Res.*, *100*(A3), 3495–3512, doi:10.1029/94JA02725.

Vautard, R., and M. Ghil (1989), Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series, *Physica D*, *35*, 395–424.

M. Ghil, D. Kondrashov, and Y. Shprits, Department of Atmospheric and Oceanic Sciences, University of California, Box 951565, 7127 Math Sciences Bldg., 405 Hilgard Ave., Los Angeles, CA, 90095-1565, USA. (dkondras@atmos.ucla.edu)