NOTES AND CORRESPONDENCE

Comments on "The Shortcomings of Nonlinear Principal Component Analysis in Identifying Circulation Regimes"

ADAM H. MONAHAN

School of Earth and Ocean Sciences, University of Victoria, Victoria, British Columbia, and Earth System Evolution Program, Canadian Institute for Advanced Research, Toronto, Ontario, Canada

JOHN C. FYFE

Canadian Centre for Climate Modelling and Analysis, Meteorological Service of Canada, and University of Victoria, Victoria, British Columbia, Canada

(Manuscript received 13 September 2005, in final form 10 February 2006)

Christiansen (2005) presents a critique of the use of nonlinear principal component analysis (NLPCA) for the detection of atmospheric circulation regimes. He first notes that for the NLPCA model architecture considered in Monahan et al. (2001, 2003) only a limited class of approximation shapes are possible, and then that the parameterization of the associated time series (and thus its probability distribution) is arbitrary up to a homeomorphism. It is further argued through numerical experiments that NLPCA of Gaussian data produces spurious multimodality, and that the analysis of 20-hPa geopotential height described in Monahan et al. (2003) is not robust. This comment addresses all three of these concerns, starting with the third.

The central point of Christiansen (2005) is that the results of Monahan et al. (2000, 2001, 2003) and Teng et al. (2004) are not robust because NLPCA will diagnose spurious regimes in data that are Gaussian (or nearly so). For this argument to hold, the analysis carried out must be identical to the analyses used in these earlier studies. In fact, the NLPCA algorithm described in Christiansen (2005) differs fundamentally from that used in the Monahan et al. and Teng et al. papers in such a way that the spurious results described are to be expected.

As is discussed in detail in Monahan (2000), a primary concern with any statistical technique is avoiding overfitting: that is, a good statistical model should be robust to the introduction of new data not used in the estimation of its parameters. As no analytic solution exists to the variational problem central to NLPCA, iterative numerical minimization techniques must be employed: a good statistical model should also be robust to the choice of initial parameter values. These two issues are discussed extensively in the context of cluster analysis by Michelangeli et al. (1995), where they are referred to respectively as the problems of reproducibility and classifiability. In the Monahan et al. and Teng et al. studies, reproducibility and classifiability were assessed using an ensemble approach. For any given dataset, NLPCA was carried out N times (where N was typically at least 50); each ensemble member used different initial parameter values and a different subset of the full dataset (generally 70%–80%, subsampled in such a way to take into account autocorrelation in the time series). The remaining ("validation") data were not used in parameter estimation, but retained to assess the skill of the resulting statistical model on "new" data. Parameter values for each ensemble member were estimated using conjugate-gradient methods with early stopping: the iterative minimization was carried out for a fixed number of iterations, or until the mean-squared error over the validation set started increasing (indicative of overfitting). Only those ensemble members for which the mean-squared error over the validation data was smaller than that over the estimation data were judged to pass the reproducibility test and therefore to

Corresponding author address: Adam H. Monahan, School of Earth and Ocean Sciences, University of Victoria, P.O. Box 3055, STN CSC, Victoria, BC V8W 3P6, Canada.
E-mail: monahana@uvic.ca

DOI: 10.1175/JCLI4002.1

be "candidate models." This algorithm does *not* look for the global minimum of the cost function, which for very rough error surfaces may be strongly affected by sampling fluctuations. Rather, it produces a representative and reproducible ensemble of models.

Each of the candidate models in the ensemble was then compared. If these models were all "close" (i.e., shared the same morphology and orientation), then it was judged that the candidate models satisfied the criterion of classifiability. If the candidate models differed considerably in shape or orientation, then it was concluded that no robust statistical model existed for the NLPCA model in its present architecture. The entire process was then repeated using a simpler NLPCA model (i.e., with one fewer "hidden neuron" in each of the compression and expansion layers). If the data contain no robust nonlinear low-dimensional structure, this algorithm will eventually converge to an NLPCA model with one hidden neuron in each of the compression and expansion layers: such a model is equivalent to classical principal component analysis (PCA). Thus, in the absence of robust low-dimensional structure in the data (in the sense of reproducibility and classifiability), NLPCA reduces to PCA.

To demonstrate that our NLPCA algorithm does not identify spurious non-Gaussian structure in Gaussian data, a random sample was drawn from a two-dimensional stochastic process $\mathbf{X} = (x_1, x_2)$ designed to have second-order statistics comparable to those of the leading two PCA modes of the Northern Hemisphere extratropical 500-hPa geopotential height analyzed in Monahan et al. (2001, 2003) [4050 points in length (45 yr of 90 winter days each), with autocorrelation e-folding time scale of 10 days and $\text{std}(x_2)/\text{std}(x_1) = 0.8$]. Nonlinear PCA with two hidden neurons in each of the compression and expansion layers was applied to this sample: out of 25 trial models, only 10 performed as well over the validation data as the training data and were considered candidate models. The candidate models fell into two classes: U-shaped and S-shaped; examples of each are displayed in Fig. 1. These two classes of candidate models have the same mean-squared errors: they are indistinguishable from the point of view of NLPCA. Their shapes are considerably different, as are the probability distribution functions of the associated (arc-length parameterized; see below) time series. The two-node NLPCA algorithm does not produce a unique, reproducible approximation for this data: we conclude that the data do not support a robust nonlinear approximation with a model of this complexity. The number of hidden neurons was then reduced to one (which reduces the algorithm to PCA), and the analysis was repeated. The results are displayed as the green

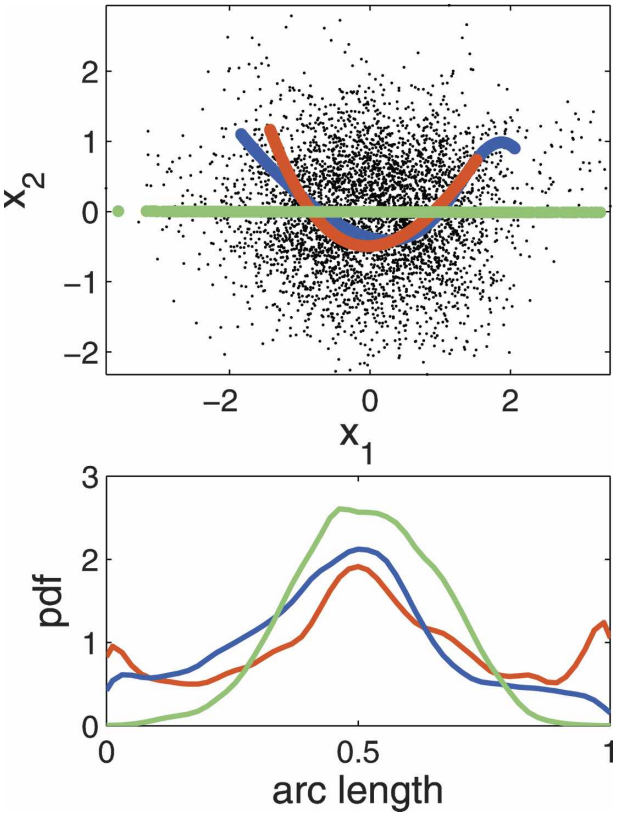

FIG. 1. (top) NLPCA approximations of Gaussian data. Spurious but nonrobust non-Gaussian structure is obtained if an NLPCA model with too many parameters is used [with different equally likely approximations exemplified by the U-shaped (red curve) and S-shaped (blue curve) candidate models]. A robust approximation (green curve) is obtained using a single hidden neuron in each of the compression and expansion layers: this is the NLPCA approximation to these data obtained using our algorithm. (bottom) Gaussian kernel estimates of the probability density functions of the arc-length parameterized time series of the NLPCA models shown in the upper panel (with corresponding color scheme).

curves in Fig. 1: the NLPCA approximation is a straight line with a unimodal time series, as should be obtained with Gaussian data. When applied carefully to a sample drawn from a Gaussian distribution, our NLPCA algorithm does not identify spurious non-Gaussian structure. A similar result is reported in Teng et al. (2006, manuscript submitted to *J. Climate*).

In the algorithm described in Christiansen (2005), while an ensemble approach is taken to address the classifiability problem, only a partial effort is made to address reproducibility (by fitting model parameters to a subset of the data). However, the algorithm described in Christiansen (2005) does not monitor the performance of the model over the validation set: there is no measure of how well the model generalizes to data not

used in the estimation of its parameters. It is to be expected that the model at the global minimum of the cost function is particularly likely to perform poorly over the validation set, and it is precisely this model that is selected for Christiansen's algorithm. Such an algorithm is highly sensitive to overfitting, and the resulting lack of robustness in the NLPCA approximations is to be expected. Without fully addressing the problem of reproducibility, the results of NLPCA (as of any nonlinear, nonparametric statistical model) are not meaningful. The spurious multimodality found in analyses of Gaussian samples is not a result of a fundamental problem with NLPCA, but rather of an insufficiently careful statistical estimation algorithm that is fundamentally different than that used in Monahan et al. (2000, 2001, 2003) and Teng et al. (2004), and is by construction highly prone to sampling errors and spurious results. We would welcome an independent repeat of the analyses of Monahan et al. (2000, 2001, 2003) and Teng et al. (2004) using a robust methodology fully taking into account both the issues of reproducibility and classifiability.

It is also important to note that the regimes diagnosed in Monahan et al. (2001, 2003) are essentially the same as those found in, for example, Cheng and Wallace (1993), Corti et al. (1999), Smyth et al. (1999), and Crommelin (2004). While NLPCA was not designed to find regimes, the regimes it diagnosed are supported by the results of completely independent studies.

As a final comment, we would like to address the issues raised in Christiansen (2005) regarding (i) the limited range of shapes available for two-neuron NLPCA approximations, and (ii) the nonuniqueness of the NLPCA time series. Regarding point (i), we do not disagree—NLPCA carried out with only two hidden neurons in the compression and expansion layers can only produce a limited range of approximation shapes. Models with more hidden neurons in these layers can produce approximations with more complex shapes, but require more data for these approximations to be robust. This is certainly a limitation of NLPCA in practical applications with datasets of finite length. Regarding point (ii), it is well known that the NLPCA time series is unique only up to an arbitrary homeomorphism (a one-to-one, onto function with a continuous inverse). From this, Christiansen (2005) concludes that the multimodality of the NLPCA approximation can be removed by an appropriate coordinate transformation. While it is indeed the case that the 1D time series characterizing the NLPCA approximation—and its probability distribution—are nonunique, the distribution of points on the full NLPCA curve in the original space [$\hat{\mathbf{X}}$ in the notation of Monahan et al. (2003)] is unique for a given statistical model. This curve characterizes the structure of the approximation in the full "physical" space, and the density of points along this curve therefore captures the "physical" density of the approximation. Any clustering along the curve is independent of how the time series is parameterized; the arc-length parameterization is a convenient parameterization because it is designed explicitly to measure the density of points along the approximation curve. The circulation regimes found in Monahan et al. (2000, 2001, 2003) cannot be removed through an appropriate change of coordinates along the NLPCA approximation.

Acknowledgments. The authors thank four anonymous reviewers for their helpful comments on an earlier draft of this manuscript. AHM is supported by the Natural Sciences and Engineering Research Council of Canada and by the Canadian Institute for Advanced Research Earth System Evolution Program.

REFERENCES

Cheng, X., and J. M. Wallace, 1993: Cluster analysis of the Northern Hemisphere wintertime 500-hPa height field: Spatial patterns. *J. Atmos. Sci.,* **50,** 2674–2696.

Christiansen, B., 2005: The shortcomings of nonlinear principal component analysis in identifying circulation regimes. *J. Climate,* **18,** 4814–4823.

Corti, S., F. Molteni, and T. Palmer, 1999: Signature of recent climate change in frequencies of natural atmospheric circulation regimes. *Nature,* **398,** 799–802.

Crommelin, D. T., 2004: Observed nondiffusive dynamics in large-scale atmospheric flow. *J. Atmos. Sci.,* **61,** 2384–2396.

Michelangeli, P.-A., R. Vautard, and B. Legras, 1995: Weather regimes: Recurrence and quasi stationarity. *J. Atmos. Sci.,* **52,** 1237–1256.

Monahan, A. H., 2000: Nonlinear principal component analysis by neural networks: Theory and application to the Lorenz system. *J. Climate,* **13,** 821–835.

——, J. C. Fyfe, and G. Flato, 2000: A regime view of Northern Hemisphere atmospheric variability and change under global warming. *Geophys. Res. Lett.,* **27,** 1139–1142.

——, L. Pandolfo, and J. C. Fyfe, 2001: The preferred structure of variability of the northern hemisphere atmospheric circulation. *Geophys. Res. Lett.,* **28,** 1019–1022.

——, J. C. Fyfe, and L. Pandolfo, 2003: The vertical structure of wintertime climate regimes of the Northern Hemisphere extratropical atmosphere. *J. Climate,* **16,** 2005–2021.

Smyth, P., K. Ide, and M. Ghil, 1999: Multiple regimes in the Northern Hemisphere height fields via mixture model clustering. *J. Atmos. Sci.,* **56,** 3704–3723.

Teng, Q., A. H. Monahan, and J. C. Fyfe, 2004: Effects of time averaging on climate regimes. *Geophys. Res. Lett.,* **31,** L22203, doi:10.1029/2004GL020840.