Beiwei Lu and William W. Hsieh *
University of British Columbia, Vancouver, British Columbia, Canada

## 1. INTRODUCTION

Principal component analysis (PCA) is widely used to extract the linear relations between variables in a dataset. To detect nonlinear relations, the nonlinear principal component analysis (NLPCA) by a 3-hidden-layer feed-forward neural network was proposed by Kramer (1991), which has been used to analyze datasets from many fields. However, the 3-hidden-layer NLPCA can be rather unstable, often resulting in the overfitting of data, especially for noisy datasets with rather few samples. Techniques such as the addition of weight penalty terms in the cost function or early stopping during training can reduce the severity of the problems, but offer no guarantee of the optimal solution. This paper shows that the instability and tendency to overfit in the 3-hidden-layer NLPCA can be well alleviated in the simplified 2-hidden-layer NLPCA.

## 2. The 3-hidden-layer NLPCA

In the 3-hidden-layer NLPCA (Fig. 1) proposed by Kramer (1991), the input signals are transfered to the 'encoding' neurons in the first hidden layer. The hyperbolic tangent function is used as the transfer function here, and again when the signal moves from the 'bottleneck' neuron in the second hidden layer to the 'decoding' neurons in the third hidden layer. Linear transfer functions are used to map from the encoding layer to the bottleneck layer and from the decoding to the output layer. Effectively, a nonlinear function maps from the higher dimension input space to the low dimensional bottleneck space, followed by an inverse transform mapping from the bottleneck space back to the original space represented by the outputs. This is an auto-associative network, in that the target and input datasets are the same, hence the cost function, which is the mean square error $\phi$ between the outputs and the inputs, is minimized by adjusting the weight and bias parameters in the transfer functions. Data compression is achieved by the bottleneck, with the single bottleneck neuron in Fig. 1 giving the leading nonlinear principal component. The numbers of encoding and decoding neurons are adjustable for the optimal fit, but are set the same for simplicity. This imposed symmetry does not affect our conclusions. The NLPCA in Fig. 1 with 3, 2, 1, 2 and 3 neurons in its 5 layers will be referred to as a 3-2-1-2-3 model.

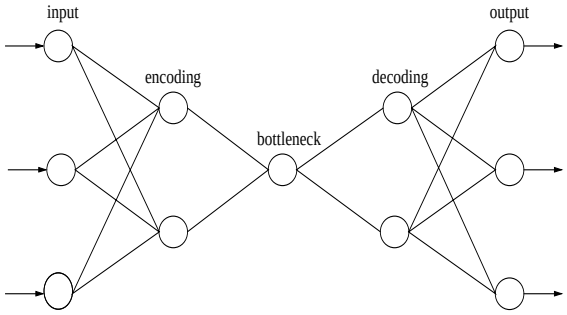


FIG. 1: A schematic diagram of the 3-hidden-layer FFNN model for performing the NLPCA. There are 3 layers of hidden neurons sandwiched between the input layer on the far left and the output layer on the far right. Next to the input layer is the encoding layer, followed by the bottleneck layer, which is then followed by the decoding layer.

It can be shown that among the derivative of $\phi$ with respect to bottleneck weights $B_m$ and bias $b$ and decoding weights $C_m$ and biases $c_m$, there are linear dependences (due to the linear transfer functions used at the bottleneck layer):

$$\sum_{m=1}^{M} B_m \frac{\partial \phi}{\partial B_m} + \sum_{m=1}^{M} C_m b \frac{\partial \phi}{\partial c_m} = \sum_{m=1}^{M} C_m \frac{\partial \phi}{\partial C_m},$$

$$\frac{\partial \phi}{\partial b} = \sum_{m=1}^{M} \frac{\partial \phi}{\partial c_m} C_m,$$

where M is the number of encoding neurons. The number of linearly independent optimization equations is 2 less than the number of weight and bias parameters. Hence the weights and biases cannot be uniquely determined and the bottleneck signal is non-unique.

As an example, the Kaplan extended sea surface temperature anomaly (Kaplan et al. 1998) from January 1856 to July 2001, covering the tropical Pacific Ocean from $22.5^o S$ to $17.5^o N$ with $5^o \times 5^o$ resolution, is analyzed. The first and second principal

* *Corresponding author address:* William W. Hsieh, Univ. of British Columbia, Dept. of Earth and Ocean Sciences, Vancouver, BC V6T 1Z4, Canada; e-mail: whsieh@eos.ubc.ca

components of the monthly anomaly, extracted using the classical principal component analysis (von Storch and Zwiers 1999), are used as the input and target datasets. The evaluated weights and biases of different runs from the same model are different. For instance, the weight of encoding neuron 1 for input series 1 $A_{11} = -1.646$ for run 1, $A_{11} = -0.174$ for run 2 while $A_{11} = -0.734$ for run 3 from the 2-2-1-2-2 model. For the run bearing the lowest MSE value of a 2-M-1-M-2 model, where M is varied from 2 to 4, Fig. 2 shows the bottleneck series $u_l$ ($l$ being the temporal index) and the output series $q_{il}$ (overlapping circles, with dots denoting the target series $p_{il}$, where $i$ indicates the ith output or target series). The bottleneck series (Figs. 2a, c and e) are non-unique. The scatterplot of the output series of the 2-2-1-2-2 run 4 forms a hump (Fig. 2b), whereas increasingly wiggly or overfitted solutions are found as M increases (Figs. 2d and 2f). In Figure 2g, the MSE values of run 1 to run 5 of the three models are shown. The MSE values are reduced but have greater standard deviations as M increases.

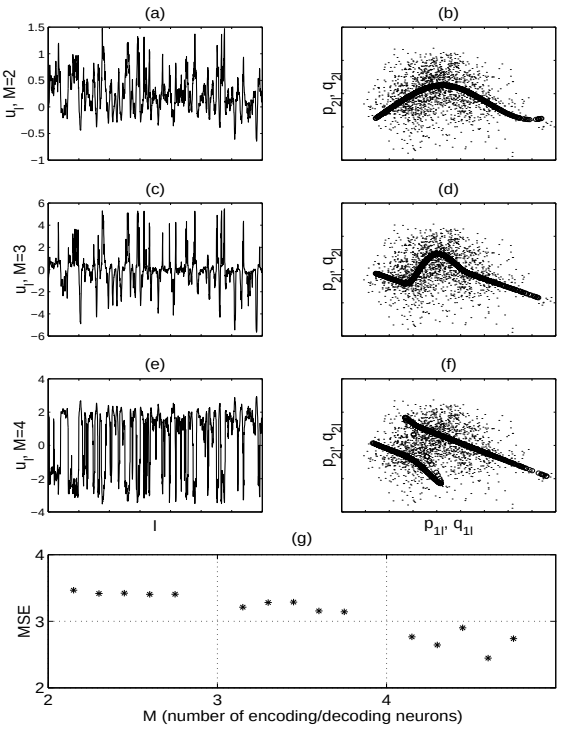


FIG. 2: The 3-hidden-layer NLPCA applied to the first and second principal components of the Kaplan extended sea surface temperature anomaly. (a), (c), (e) Bottleneck and (b), (d), (f) output series (overlapping circles) of 2-2-1-2-2 run 4, 2-3-1-3-2 run 5, 2-4-1-4-2 run 4. (g) MSE values of run 1 to run 5 from the 2-M-1-M-2 model, where M is varied from 2 to 4. The target data are shown as dots in (b), (d) and (f).

## 3. The 2-hidden-layer NLPCA

We will show that by eliminating the encoding layer and replacing the linear function at the bottleneck layer by a nonlinear one, the overfitting and non-uniqueness problems are alleviated. Figure 3 shows the 2-hidden-layer NLPCA with three input, one bottleneck, two decoding and three output neurons, i.e, a 3-1-2-3 model. The hyperbolic tangent function is used at both hidden layers while the linear transfer function is retained at the output layer.

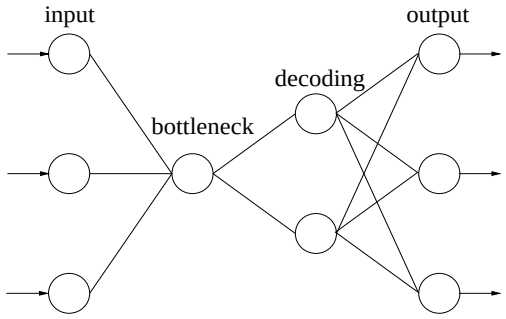


FIG. 3: A 2-hidden-layer NLPCA. From left to right are the input, bottleneck, decoding and output layers of neurons.

A dataset of 600 sampling points is generated from the Lorenz attractor, a 3-component chaotic system (Lorenz 1963), and analyzed by the 3-1-M-3 model, where M is varied from 4 to 14. The MSE values group into two, one is around a local minimum near 14.1 and the other around the global minimum below 14.0. Figure 4a shows the MSE values of 20 runs for each M, as M increases. To see the lowest MSE value of each model, Figure 4b shows the MSE values between 13.96 and 13.99. The lowest MSE value of all the models is 13.97 attained for M=6. To see whether the solutions of the M=7 and 8 models are the same as the M=6 model, Figure 4 shows the target series $p_{il}$ (thin line) and the output series $q_{il}$ (dots), $p_{1l}$ and $q_{1l}$ in (c), $p_{2l}$ and $q_{2l}$ in (d) and $p_{3l}$ and $q_{3l}$ in (e), of the runs bearing MSE$<$ 13.99 from the 3-1-6,7,8-3 models (solid circles in Fig. 4b). The output series of the runs included are indistinguishable from one another and smoothly approximate the target series. The overfitting problem is well restrained. The corresponding bottleneck series $u_l$ (Fig. 5a) displays signs which are often opposite among different runs giving rise to symmetric patterns. In general, $u_l$ reverses sign at around $l$ = 260 and $l$ = 480. The major cause of the deviations among the bottleneck series seems to be the inclusion of bottleneck bias, as we will test by deleting the bottleneck bias from the network.

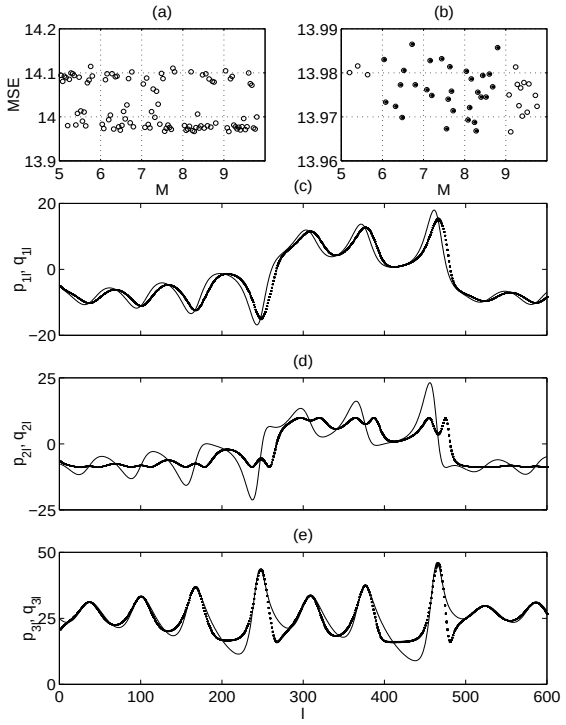


FIG. 4: The 2-hidden-layer NLPCA applied to the Lorenz dataset. (a) MSE values and (b) MSE values between 13.96 and 13.99, from the 3-1-M-3 model, where M is varied from 5 to 9. (c), (d) and (e) The Lorenz data (thin line) and the output series (dots) of the runs with MSE¡13.99 from the 3-1-6,7,8-3 models.

The Lorenz dataset is again analyzed by the 3-1$^b$-M-3 model, without the bottleneck bias as denoted by the superscript $b$, where M is again varied from 4 to 14 with 20 runs for each M. The resultant MSE values (not shown) suggest the same two groups as before and the output series of the runs with MSE $\leq$ 13.99 from the 3-1 $^b$ -6,7,8-3 models (not shown) strongly resemble those from the 3-1-6,7,8-3 models (Fig. 4a,c-e). Again, the lowest MSE of the new models is 13.96$-$13.97 for $M \geq 6$. The scatter of $u_l$ from 3-1$^b$-6,7,8-3 model runs with MSE $\leq$ 13.99 (Fig. 5b) is much reduced compared to the 3-1-6,7,8-3 model results (Fig. 5a). Hence, eliminating the bottleneck bias parameter significantly reduces the deviation of the bottleneck series for different runs.

Both with MSE of 13.97, run 6 and run 13 of the 3-1$^b$-6-3 model have indistinguishable output series and indistinguishable bottleneck series. However, the output biases for run 6 and run 13, respectively, are $d_1$ = 0.6370 and 1.0214, $d_2$ = $-$0.5604 and 0.2982, $d_3$ = 5.2902 and 3.3330 ($d_i$ is the bias of output neuron i). This suggests testing the NLPCA without the output bias parameters. The Lorenz dataset is analyzed using the 3-1$^b$-M-3$^b$ model, i.e., without bottleneck and output biases as denoted by the superscript b, where M is still varied from 4 to 14 with 20 runs for each M. Again the MSE values (not shown) fall into two groups as for the 3-1-M-3 model (Fig. 4a) and the lowest MSE value of the 3-1$^b$-M-3$^b$ model is 13.97, attained for M$\geq$6.

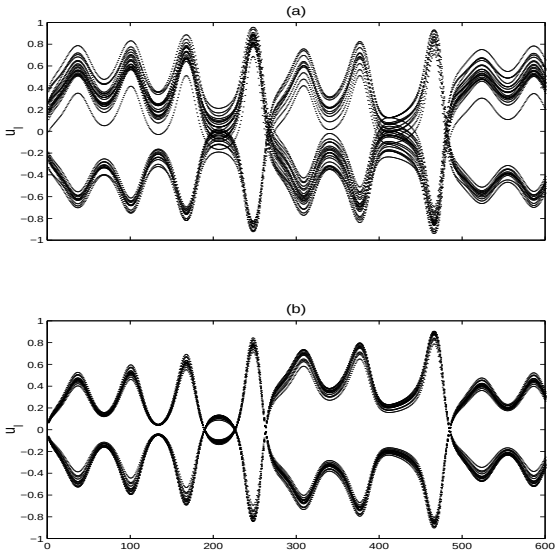


FIG. 5: The bottleneck series $u$ of the runs with MSE¡13.99, (a) from the 3-1-6,7,8-3 models and (b) from the 3-1$^b$-6,7,8-3 models.

The 3-1$^b$-6-3$^b$ model, with the least number of parameters among models with similar low MSE values, is optimal according to the principal of parsimony (Burnham and Anderson 1998) and its run 15, bearing the lowest MSE value, offers the optimal solution. Figure 6 shows the optimal solution (Figs. 6a, b, c) and a suboptimal solution from the 3-1$^b$-5-3$^b$ model with MSE=14.89 (Figs. 6d, e, f). The suboptimal solution shows a linear relation between $q_{1l}$ and $q_{2l}$ and a less curved relation between $q_{1l}$ and $q_{3l}$ than in the optimal solution.

As a second example, the first two principal components of the Kaplan extended sea surface temperature anomaly are analyzed by the 2-1$^b$-M-2$^b$ model for M=3,4,5. The MSE values steadily converge to near 3.5 (Fig. 7a). The bottleneck series and the output series of the runs bearing the lowest MSE from M=3,4,5 models are very similar, and the results of the 2-1$^b$-5-2$^b$ model run 1 are shown in Fig. 7. The non-uniqueness and overfitting problems are well restrained in the simplified 2-hidden-layer NLPCA, the 2-hidden-layer NLPCA with neither bottleneck nor output biases.

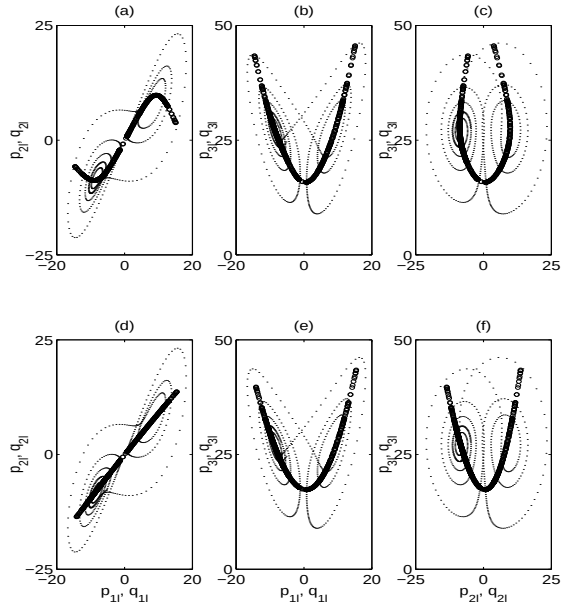


FIG. 6: (a), (b) and (c) The output series (circles) by the 3-1$^b$-6-3$^b$ NLPCA model with MSE=13.97 and the original Lorenz data (dots). (d), (e) and (f) The output series by a 3-1$^b$-5-3$^b$ suboptimal solution (MSE=14.89). (a) and (d) $q_{1l}, p_{1l}$ versus $q_{2l}, p_{2l}$. (b) and (e) $q_{1l}, p_{1l}$ versus $q_{2l}, p_{2l}$. (c) and (f) $q_{2l}, p_{2l}$ versus $q_{3l}, p_{3l}$.

## 4. The simplified 2-hidden-layer NLPCA with two bottleneck neurons

As the two loops of the Lorenz attractor are not represented by NLPCA with one bottleneck neuron, the Lorenz dataset is then analyzed using the 3-2$^b$-M-3$^b$ model, where M is varied from 20 to 30. Figure 8a shows the MSE values ($\leq$15) of run 1 to run 10 for each M. One group of MSE values is around a local minimum about 14, while the other group is around the global minimum under 1. Figure 8b shows the MSE values between 0.6 and 0.8. The lowest MSE value of all the models is near 0.6, attained by the 3-2$^b$-22-3$^b$ model run 7 and the 3-2$^b$-29-3$^b$ model run 3 (darkened circles in Fig. 8b). Figure 8c-e shows the target series (thin line) and the output series (dots) of these two runs, where $q_{1l}$ and $q_{3l}$ are indistinguishable from $p_{1l}$ and $p_{3l}$ respectively, while $q_{2l}$ is often indistinguishable from $p_{2l}$. The scatterplots of the bottleneck series, $u_{1l}$ versus $u_{2l}$ of the 3-2$^b$-22-3$^b$ run 7 (Fig. 9a, circles) and $-u_{1l}$ versus $u_{2l}$ of the 3-2$^b$-29-3$^b$ run 3 (Fig. 9a, crosses) are similar, showing that the non-uniqueness of the bottleneck series is not serious. Figure 9b-d shows the scatterplots of the output series of the 3-2$^b$-22-3$^b$ run 7, where intersecting curves are well simulated.

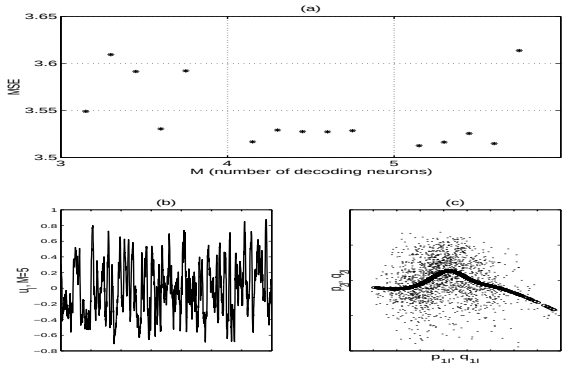


FIG. 7: The simplified 2-hidden-layer NLPCA applied to the first two principal components of Kaplan extended sea surface temperature anomaly. (a) MSE values from the 2-1$^b$-3,4,5-2$^b$ models. (b) Bottleneck and (c) output series of the 2-1$^b$-5-2$^b$ run 1.

## 5. Representation of the zonal wind in the equatorial stratosphere

The zonal winds in the tropical stratosphere exhibit a predominant quasi-biennial oscillation (QBO) (Baldwin et al. 2001). Because the reconstruction of the QBO wind by the leading two principal components of the principal component analysis (Wallace et al. 1993) on the height-time record and a linear construction of composite QBO cycle lose some well-known features (Naujokat, 1986; Baldwin et al. 2001), many authors used the wind at one arbitrarily-chosen level in their studies to characterize the phase of the QBO. Recently, using the 3-hidden-layer NLPCA with a circular bottleneck neuron (Hamilton and Hsieh, 2002), the zonal winds are better modelled than the linear analyses. However, the circular bottleneck neuron is limited to only carrying phase information. So the zonal winds are reanalysed here using the simplified 2-hidden-layer NLPCA with two bottleneck neurons. The data are from the monthly means of the zonal wind component measured twice-per-day by balloons above Canton Island (2.8$^o$N) during January 1956 through August 1967, Gan (0.7$^o$S) from September 1967 through December 1975, and Singapore (1.4$^o$N) from January 1976 through December 2000 (Marquardt and Naujokat 1997). Values at 70, 50, 40, 30, 20, 15 and 10 hPa (i.e. from about 20 km to 30 km altitude), with the 45-year mean removed but the weak seasonal cycle retained, are used and hereafter called the QBO wind for convenience.

The QBO wind is analyzed using the 7-2$^b$-M-7$^b$ model, where M is varied from 10 to 30 with 20 runs for each M. The resultant MSE values (not shown) suggest the global minimum around 9, was

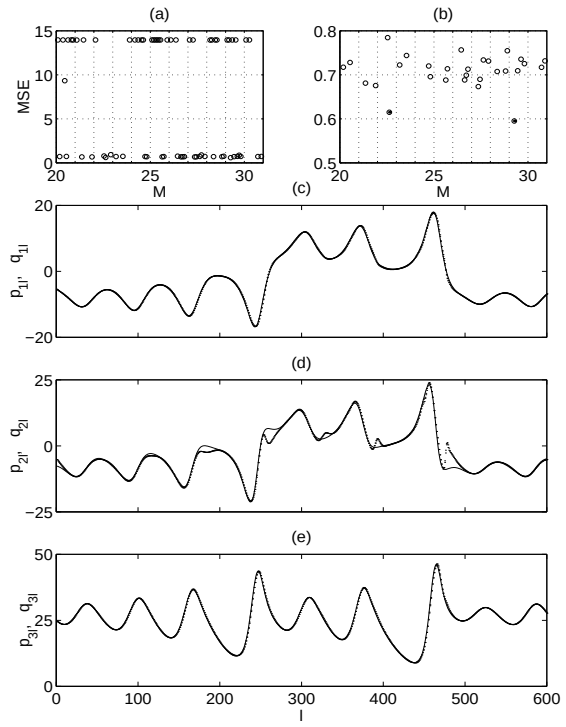


FIG. 8: The simplified 2-hidden-layer NLPCA applied to the Lorenz dataset. (a) MSE values and (b) MSE values between 0.96 and 0.99 from the 3-$2^b$-M-$3^b$ model, where M is varied from 20 to 30. (c), (d) and (e) The Lorenz data (thin line) and the output series (dots) of the 3-$2^b$-22-$3^b$ model run 7 and 3-$2^b$-29-$3^b$ run 3.

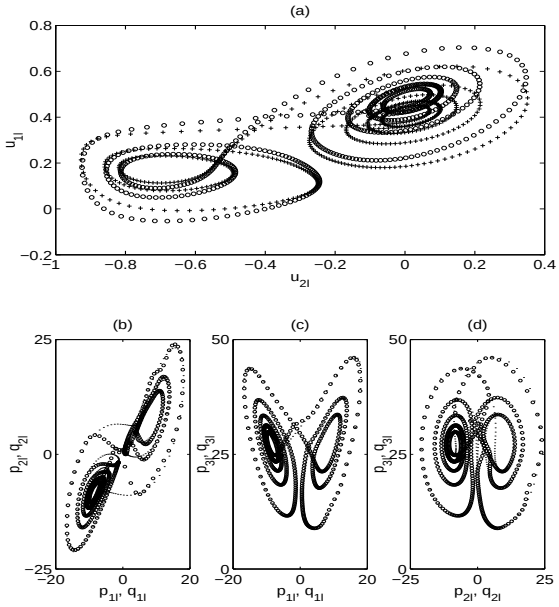


FIG. 9: (a) The scatterplots of the bottleneck series of the 3-$2^b$-22-$3^b$ model run 7 (circles) and the 3-$2^b$-29-$3^b$ run 3 (crosses). (b), (c) and (d) The scatterplots of the output series of 3-$2^b$-22-$3^b$ run 7 and the Lorenz data. (b) $q_{1l}, p_{1l}$ versus $q_{2l}, p_{2l}$. (c) $q_{1l}, p_{1l}$ versus $q_{3l}, p_{3l}$. (d) $q_{2l}, p_{2l}$ versus $q_{3l}, p_{3l}$.

attained by about half of the runs. The optimal solution with the lowest MSE value of 9.0 is from the 7-$2^b$-27-$7^b$ model. At the 7 levels from 70 to 10 hPa, the correlation of the output and target series are 0.867, 0.969, 0.986, 0.993, 0.991, 0.990 and 0.977, respectively, and the root mean square errors (RMSE) are 3.26, 3.23, 2.61, 2.07, 2.56, 2.85 and 3.99. The averaged correlation and RMSE of the seven layers are 0.968 and 2.94, superior to those from the reconstruction by the leading two principal components of 0.945 and 4.25 (Wallace et al. 1993) and by 3-hidden-layer NLPCA with a circular bottleneck neuron of 0.957 and 3.73 (Hamilton and Hsieh, 2002). The two bottleneck series are out of phase, varying in both amplitude (Fig. 10a) and phase (Fig. 10b). The variations of the phase show cycles of 23-35 months in length (Fig. 10b).

## 6. Conclusions

This study has found that the non-uniqueness and overfitting problems in the 3-hidden-layer NLPCA, an under-determined model, are well alleviated by the simplified 2-hidden-layer NLPCA, as demonstrated by different datasets - the sea surface temperature anomaly over the tropical Pacific Ocean, the Lorenz chaotic system and the QBO wind.

## 7. References

**Baldwin,** M., L. Gray, T. Dunkerton, K. Hamilton, P. Haynes, W. Randel, J. Holton, M. Alexander, I. Hirota, T. Horinouchi, D. Jones, J. Kinnersley, C. Marquardt, K. Sato and M. Takahashi, 2001: The Quasi-biennial oscillation. *Rev. Geophys., 39*, 179-229.

**Burnham,** K. P. & D. R. Anderson, 1998: *Model selection and inference, a practical information-theoretic approach*. New York: Springer, 354 pp.

**Hamilton,** K. and W.W. Hsieh, 2002: Representation of the QBO in the tropical stratospheric wind by nonlinear principal component analysis. *J. Geophys. Res*. 107(D15), DOI: 10.1029/2001JD001250.

**Kaplan,** A., M. Cane, Y. Kushnir, A. Clement, M. Blumenthal & B. Rajagopalan, 1998: Analyses of global sea surface temperature 1856-1991. *Journal of Geophysical Research*, **103,** 18,567-18,589.

**Kramer,** M. A., 1991: Nonlinear principal component analysis using auto-associative neural networks. *American Institute of Chemical Engineers Journal*, **37,** 233-243.

**Lorenz,** E. N., 1963: Deterministic non-periodic flow. *Journal of the Atmospheric sciences*, **20**, 130-141.

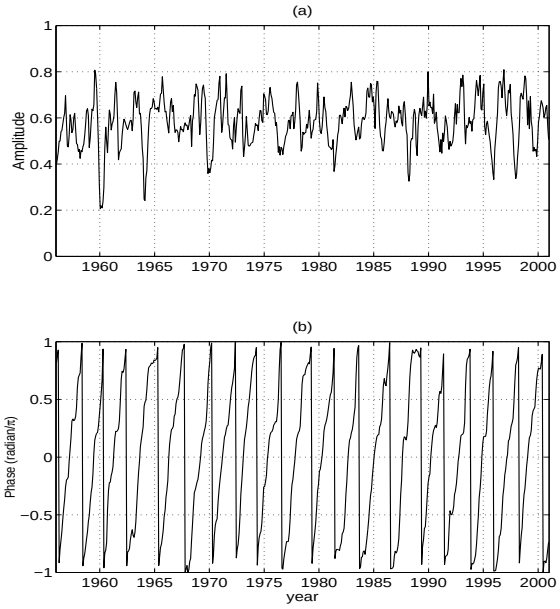


FIG. 10: (a) The amplitude and (b) the phase of the optimal QBO wind bottleneck series.

**Marquardt,** C. and B. Naujokat, 1997: An update of the equatorial QBO and its variability. *World Meteorological Organization Technical Document 814*, 87-90.

**von** Storch, H. & F. W. Zwiers, 1999: *Statistical analysis in climate research*. Cambridge University Press, 494 pp.

**Wallace,** J. M., L. Panetta and J. Estberg, 1993: A phase-space representation of the equatorial stratospheric quasi-biennial oscillation. *J. Atmos. Sci. 50*, 1751-1762.