Article ID: 1006-8775(1999) 02-0189-10

A GENERALIZED CANONICAL MIXED REGRESSION MODEL FOR ENSO PREDICTION WITH ITS EXPERIMENT*

JIANG Zhi-hong (江志红)[1], DING Yu-guo (丁裕国)[1] and ZAI Pan-mao (翟盘茂)[2]

(1.*Nanjing Institute of Meteorology, Nanjing* 210044 *China*; 2. *National Climate Center of China, Beijing* 100081 *China*)

ABSTRACT: A scheme is proposed for predicting NINO-region SST in terms of a generalized canonical mixed regression model based on principal component canonical correlation analysis (PC-CCA). and into the scheme are introduced such techniques as EEOF, PRESS criterion and consensus prediction. By optimizing physical factors and selecting optimal model parameters, experiments were made successful in predicting the NINO SST index for 1 to 4 seasons to follow. The scheme is shown to be stable in operation and its total technical level compares well with that of the model published in NOAA/NWS/NCEP CPC *Climate Diagnostics Bulletin*, but the number of factors needed in our scheme is much fewer than that for the CPC's model in dealing with the same problems. This makes it possible to establish an operational ENSO monitoring system in China.

Key words: ENSO prediction; canonical regression; prediction scheme

CLC number: P456.3 **Document code:** A

I. INTRODUCTION

It is well known that nowadays the ENSO phenomenon represents the strongest signal detected of the global climate system; its genesis and evolution have been the great concern of meteorologists and realization of ENSO prediction on a global basis is the key link in exploring the ENSO-related climate behaviors. ENSO prediction, however, is both a heated issue and a hard nut in the context of short-term climate prognosis. Despite the fact that ENSO genesis and development show a range of signs, the variation in SST anomaly is the immediate indicator. While prediction of SSTA over some key regions in the equatorial Pacific (Zebiak and Cane, 1987) by means of simple and complicated air-sea coupling models has been successful since the mid 1980s there remain many deficits, e.g., season-varying drift of product's stability, less steadiness at different time lead and substantial difference in extended lengths between different models — all these await further improvement, and particularly, prediction for > 6 months in advance remains difficult.

Statistical prediction models are a less costly and more efficient approach to climate prediction. After detailed consideration of physical mechanisms or total features of some aspects of the climate system, the application of an appropriate statistical prediction model and scheme is likely to reach the goal. Studies show that for ENSO prediction the use of such statistical models as canonical correlation analysis (CCA), principal oscillation pattern (POP), linear inverse modeling

* **Received date:** 1998-09-08; **revised date:** 1998-12-07

Foundation item: National Science &Technology Scaling Project of China in the 9th Five-Year-Plan period (96-908-04-02-4)

Biography: JIANG Zhi-hong (1963 –), female, native from Wuxi City Jiangsu Province, associate professor and Ph.D. holder at Nanjing Institute of Meteorology, undertaking the study of climatology.

(LIM) and singular spectral analysis (SSA) has resulted in equally good output (Barnston and Ropelewski, 1992; Xu and Storch, 1990; Penland and Magoian, 1993). And preliminary results have been attained at home and abroad and put into operation in some countries. The writers developed a canonical autoregression prediction model for meteorological element fields (Ding et al., 1996) and now propose thereupon a generalized canonical mixed regression model with experiments thereby done on NINO SST index. To verify the scheme we shall present experimental results in comparison to predictions. This will be contribution to the development of an ENSO monitoring system in the 9th Five-Year-Plan period in China.

II. MODEL AND PREDICTIVE SCHEME

1. *Model*

Set $Y=(y_1, y_2, \cdots, y_q)'$ and $X=(x_1, x_2, \cdots, x_p)'$ to be the field of predictive variables (predictands) and predictors, respectively, where $y_i = y_i(t)$ and $x_i = x_i(t)$ with $i = 1.2.\cdots, q$ and p, respectively, and $t = 1, 2, \ldots, n$, in which q, p and n stand, in order, for the number of predictands, predictors and length of series. Without the loss of generalization, we assume that the variables in X and Y are mathematically made mean-nullified.

To extract dominant signals, principal component analysis (PCA) is undertaken of X and Y fields, leading to

$$\begin{cases} \underset{(p\times n)}{a\ (t)} = \underset{(p\times p)}{E'}\ \underset{(p\times n)}{X} \\ \\ \underset{(q\times n)}{\beta\ (t)} = \underset{(q\times q)}{F'}\ \underset{(q\times n)}{Y} \end{cases} \tag{1}$$

where $a_t = [a_1(t), a_2 \cdots, a_p(t)]$ and $\beta_t = [\beta_1(t), \beta_2 \cdots, \beta_q(t)]$ represent, respectively, the standardized principal components (in vectorial form, the same below) of X and Y, $E = e_1, e_2, \cdots, e_p$ and $F = (f_1, f_2, \ldots, f_q)$ the load vectors of the related principal components.

If, under an appropriate criterion, we take the first $p_1 < p$ $(q_1 < q)$ principal components as dominant signals of the predictors (predictands), then we shall obtain, through CCA of the newly-formed variables $a_t = [a_1(t), a_2 \cdots, a_{p_1}(t)]$ and $\beta_t = [\beta_1(t), \beta_2 \cdots, \beta_{q_1}(t)]$, the canonical correlation variables, corresponding to associated canonical correlation coefficients $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots \lambda_{q^*})$, with $q^* \leqslant \min(p_1, q_1)$. Consequently,

$$\begin{cases} U(t) = \Gamma'\ a\ (t) \\ \\ V(t) = \Phi'\ \beta\ (t) \end{cases} \tag{2}$$

in which $\Gamma = (\gamma_1, \cdots \gamma_{p1})'$ and $\Phi = (\varphi_1, \cdots \varphi_{q1})'$ denote the vectors of corresponding weighted coefficients. From Eqs.(1) ~ (2) we are allowed to approximately describe the sequences of the original fields of predictors and predictands by virtue of linear combinations of canonical correlation variables.

$$X \longrightarrow \tilde{X} = E(\Gamma')^{-1}U(t) \tag{3}$$

$$Y \longrightarrow \tilde{Y} = F(\Phi')^{-1}V(t) \tag{4}$$

Allowing for orthogonality of such variables, we find

$$\tilde{Y} = HV(t) \tag{5}$$

with

$$H = YV'(t) \tag{6}$$

Following regression theory we can derive from the above relations a statistical prognostic expression for calculating the leading signals of the predictand fields in terms of those of the predictor fields, viz.,

$$\tilde{Y} = H\Lambda U(t) = H\Lambda\Gamma' E'X = BX \tag{7}$$

where $B = H\Lambda\Gamma' E'$ is a matrix of regression coefficients, of which the estimate depends on the canonical correlation coefficient Λ and the number of principal components, p_1 and q_1, taken from X and Y. Evidently, the statistical model is PC-CCA based canonical regression model.

2. *Predictive scheme*

The scheme consists of the following.

a) The prognostic fields are generated by consecutive three-month running mean SST index series, separately, from NINOs 1+2, 3, 4 and 3.4 in the equatorial eastern Pacific for January 1951 ~ December 1997.

b) The predictor fields are based on previous Southern Oscillation index (SOI), Tahiti surface pressure (TAHITI), Darwin ground pressure (DARWIN), equatorial western Pacific 850-hPa zonal wind index (WPZW), the central Pacific analog (CPZW), the eastern Pacific (EPZW), the equatorial central eastern Pacific OLR index and equatorial Pacific 200-hPa zonal wind (ZW), together with the previous SSTA — all the series range from January 1979 to December 1997 but those of SOI, TAHITI and DARWIN that have the same length as the SST series.

c) Following (7), techniques of consensus forecasting and EEOF are introduced to augment the volume of prognostic information as much as possible and include the quasi-periodic evolution at greater than interannual scales in ENSO cycles. The basic field sequences of predictors (X_1、X_2、X_3、X_4) are generated by the element fields for predictions 1 to 4 seasons in advance (each based on 3-month running means).

d) The scheme is constructed for predicting the SST for 1, 2, 3, and 4 seasons in advance. For example, given a series of the SSTA for JFM (the first season) of a particular year, the mean SST indices for the coming AMJ over the four sea regions aforementioned are to be predicted.

Then we assume the given SST index (JFM) to be (X_1), that of the previous October-December (X_2), of the previous July-September (X_3) and of the previous April-June (X_4) (X_1 to X_4 constituting a predictor series) and we thus formulate a predictive scheme for 1 season leading predictions.

e) Other factors are put into the prognostic field in accordance with the related season. Since the 4 NINO-region SSTA are involved synchronously in constructing the predictor and predictand fields, the number of primitive variables in making the model ought to be 4 times the number of the NINO regions, to which added are other factors.

f) The sample length in establishing the model has effect on predictive accuracy. For convenience, all the samples were sliding 30 year running means. As we know, such information is subject to time-dependent attenuation in transmission, and also, in order to compare the model ability, *the use of n=30 (years) as the length is understandable despite the fact that it is*, to some degree, an artifact (for details refer to a later part).

III. NUMERICAL EXPERIMENT ON THE USEFULNESS OF THE MODEL AND ITS SCHEME

1. *Experiment with an independent sample under PRESS criterion*

PRESS criterion is widely used in regression analysis of modern times. Compared to Residual Square Sum (RSS) criterion, PRESS shows great merits in model identification, parameter choice and predictive skill assessment so that we shall thereby select optimal parameters and appraise the model ability (Yao et al., 1992).

Let the r-th specimen (experimental point) be removed, in order, from the samples X and Y, and construct a model by virtue of new samples consisting, separately, of the remaining n-1 specimens. The original model

$$\tilde{Y} = BX \tag{8}$$

which is denoted as

$$\tilde{Y}(t) = B(r)X(r) \tag{9}$$

where B (r) signifies a matrix of estimate coefficients, resulting from the removal of the r-th specimen out of the preditand field $Y=(y_1^{(r)},\cdots,y_q^{(r)})'$ and of the predictor field $X=(x_1^{(r)},\cdots,x_p^{(r)})'$. And if all predictors at r-th time be put into Eq.(9), then we get the estimates of the predictands at this time. Therefore, if from the first specimen (at r-th time) we start to build a model using the same procedure, in order, then we will obtain the predicted values of all the specimens (predictions for all the time intervals). Next, we shall find the correlation coefficients between the predictions of all subsets of different model parameters (p_1, q_1, q^*) and actual measurements for all the specimens, and select model (7) satisfied by the parameter subset $(p_1, q_1, q^*)_{R=max}$ (correlation coefficients being maximum) as the optimal prognostic model.

Tab.1 presents the parameter-varying variation in December ~ February mean prediction

made one season in advance over the NINO regions. It is evident that the prediction is optimal at p_1 =5, q_1= 2 or 1 and q^*= 1, but their increase in magnitude would lead to accuracy reduction, indicating that a particular combination of the parameters is possibly responsible for maximum signal-to-noise ratio in the model (i.e., optimal prediction). Obviously, for different predictive time periods, seasons and combination of factors, optimal parameter subset (p_1 , q_1 , q^*) may differ. Our experiments demonstrate that such prediction would be optimal for combinations of p_1 =5 ~ 7, q_1 =1 ~ 2 and q^*=1 ~ 2 in general.

Tab.1 Model parameters-dependent variation in December-February mean prediction for the NINOs made one season in advance

	q_1 =1	q_1 =2		q_1 =3		
	q_1 =1	q^* =1	q^* =2	q^* =1	q^* =2	q^* =3
p_1 =3	0. 760	0. 769	0. 754	0. 768	0. 751	0. 741
p_1 =4	0. 770	0. 774	0. 760	0. 779	0. 768	0. 758
p_1 =5	0. 794	0. 794	0. 788	0. 790	0. 782	0. 786
p_1 =6	0. 792	0. 785	0. 786	0. 787	0. 771	0. 782
p_1 =7	0. 781	0. 778	0. 768	0. 780	0. 757	0. 765

2. *Choice of factors and their combinations for experiment*

Because, except SSTA, SOI, TAHITI and DARWIN, the other five factors have data starting from January, 1979, we adopt the combinations for experiments based on different sample length.

(1) Experiment I

Only SOI and previous NINO SST index were employed to make a predictor field, separately, for predicting SST index 1 to 4 seasons (based on 3-month running mean) in advance. Obviously, the experiments conducted under PRESS criterion with an independent sample allowed to compare the predictions with SOI included to those without. Tab.2 summarizes the results. It follows that, on average, the predictions with the inclusion of SOI are significantly improved, regardless of the predictive time lead, especially those for 2 and 4 seasons in advance, resulting in the improvement of correlation coefficients, on average, by approximately 0.1. Such experiments are especially useful for prediction of April-June (early summer) and May-July (mid summer) SST index in advance that has been poorly predicted before and, with prolonged predictive length of time, the improvement is noticeable and even in excess of 20% in some cases.

(2) Experiment II

We give the comparison of predictions with different predictors included (for the acronyms see subsection II .2) for predicting 3 seasons in advance (Tab.3).

Tab.2 Comparison of predictions with SSTA as the only predictor and SOI included and without 1 to 4 seasons in advance for NINO3.4 (the independent sample length n = 45 for 1953-1997)

Months	1 season leading		2 seasons leading		3 seasons leading		4 seasons leading	
	SST only	SST SOI	SST only	SST SOI	SST only	SST SOI	SST only	SST SOI
12-2	0.91	0.93	0.89	0.87	0.68	0.67	0.38	0.36
1-3	0.87	0.92	0.85	0.85	0.72	0.72	0.50	0.49
2-4	0.84	0.85	0.76	0.80	0.70	0.69	0.54	0.54
3-5	0.83	0.85	0.59	0.66	0.61	0.60	0.42	0.44
4-6	0.84	0.85	0.47	0.64	0.38	0.54	0.39	0.45
5-7	0.69	0.67	0.37	0.47	0.21	0.34	0.21	0.29
6-8	0.63	0.71	0.33	0.37	0.15	0.20	0.15	0.16
7-9	0.76	0.83	0.50	0.50	0.10	0.44	0.06	0.35
8-10	0.85	0.90	0.58	0.58	0.32	0.36	0.11	0.29
9-11	0.85	0.91	0.70	0.75	0.34	0.33	0.12	0.24
10-12	0.90	0.94	0.85	0.87	0.55	0.54	0.28	0.49
11-1	0.93	0.96	0.90	0.85	0.62	0.62	0.45	0.46
mean	0.82	0.86	0.60	0.69	0.44	0.50	0.30	0.38

In the context of January 1979-December 1997 sample size, experiments were performed in terms of a range of combinations of different predictors. Evidence suggests that despite the limited length there differ the effects on the prognostic model of combinations of different factors having varied physical implication, with the results given in Tab.3. It is seen therefrom that with such factors of clear physics as SOI and CPZW included, the predictions are greatly improved for the NINOs both for the coming summer and winter, and particularly, as SOI and CPZW are both introduced, correlation coefficients get improved more remarkably than those just from the SST field (refer to Tab.3 in comparison to Tab.2 where the runs with such combinations reach the maximum). This clearly demonstrates, from one angle, that the ENSO physical mechanism bears an intimate relation to a range of forcings from interannual oscillation of tropical air-sea interaction. However, it is evident that the model-establishing sample length and experimental sample size have innegligible influence on predictions. For instance, in contrast to Tab.3 built on

Tab.3 Comparison of predictions with different factors involved for 3 seasons in advance for 3 seasons in advance for NINOs (the independent sample size in 1981-1997)

Factor	Prediction for Dec.~Feb.			Prediction for May~July		
	NINO3	NINO4	NINO3.4	NINO3	NINO4	NINO3.4
SST only	0.372	0.743	0.535	0.004	0.422	0.126
SST,SOI	0.402	0.743	0.543	0.016	0.525	0.260
SST, CPZW	0.388	0.726	0.544	0.044	0.331	0.094
SST, EPZW	0.409	0.715	0.527	0.002	0.485	0.186
SST, WPZW	0.239	0.633	0.420	-0.012	0.406	0.098
SST, ZW	0.329	0.707	0.493	-0.008	0.387	0.095
SST, OLR	0.289	0.642	0.447	0.085	0.373	0.065
SST.SOI CPZW	0.467	0.775	0.610	0.108	0.602	0.350
SST,SOI EPZW	0.417	0.723	0.532	0.020	0.581	0.282
SST,SOI CPZW, EPZW	0.464	0.749	0.575	0.040	0.574	0.313

a 16-year sample for model-establishing to examine an independent sample for 17 years, Tab.2 is based a 30-year sample size for the establishment to examine the counterpart for 45 years, resulting in more reliable predictions. Results given in Tab.3, however, are just for comparison of relative optimum among the combinations.

3. *Prediction skill tests*

Based on results from Exps.I and II, a prognostic scheme was constructed to compare predictions for the NINOs and seasons. The previous SOI and SST index were utilized as the predictor fields and the model parameters optimized following PRESS criterion with the model-establishing sample length $n = 30$ years. The predictions for the NINOs and seasons are given for comparison.

1) COMPARISON OF PREDICTIONS FOR THE DIFFERENT NINO REGIONS

Fig.1 depicts the curves of independent sample prognosis versus measurements (in anomaly form) in 1953-98 for NINO3.4. Tab.4 reveals that the more the seasons in advance, the poorer the predictions. On the whole, however, averaged predictions 1 ~ 4 seasons leading have higher correlation coefficients. For any of the NINOs, the skill of our predictions more than one season in advance is always higher compared to that of persistence prediction. In comparison, the predictions are optimal for NINOs 3.4 and 4, leading to the correlations of 0.85, 0.70, 0.50 and 0.35, respectively, for 1, 2, 3 and 4 season leading predictions in striking contrast to prediction for NINO1+2. In addition, we prepared a diagram of 1-3 season leading predictions versus observations for each of the NINOs after 1980 (figure not shown), where the 2-season leading predic-

Tab. 4 Correlations of 1-4 season leading predictions and persistence forecasting (bracketed) to observations for each of the NINOs*

Season leading	NINO 1+2	NINO3	NINO4	NINO 3.4
1	0.73 (0.72)	0.84 (0.74)	0.85 (0.83)	0.87 (0.78)
2	0.42 (0.35)	0.66 (0.37)	0.70 (0.57)	0.72 (0.42)
3	0.17 (0.04)	0.42 (0.04)	0.50 (0.33)	0.50 (0.09)
4	0.10 (-0.12)	0.29 (-0.13)	0.35 (0.11)	0.37 (-0.10)

*The used independent variation covers 1953-1997 (n=45)

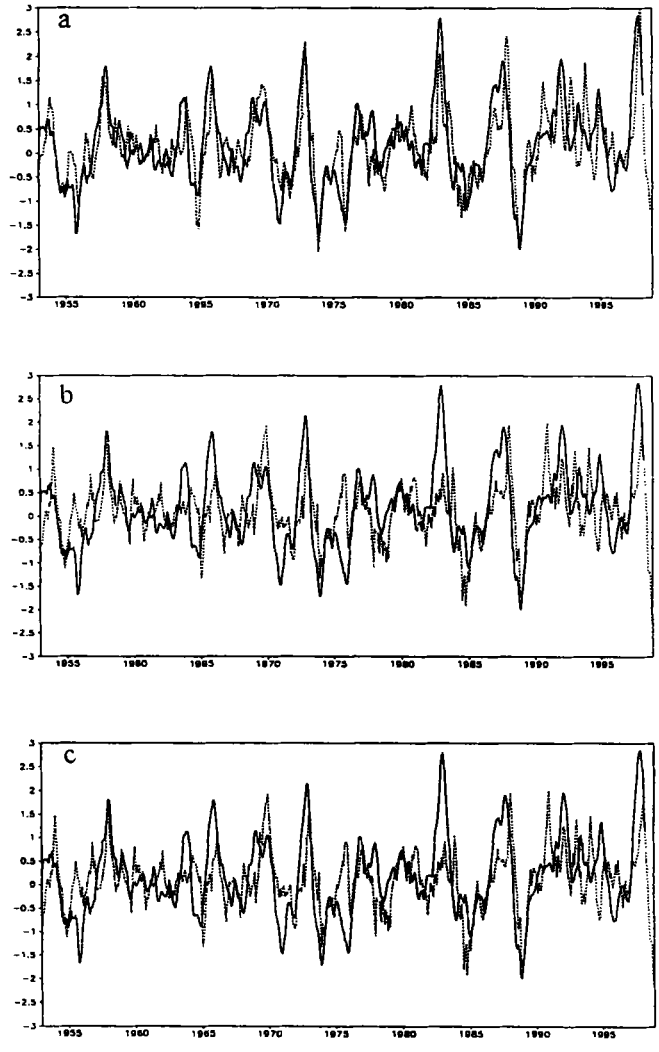

Fig.1 NINO-3.4 SST anomaly predictions (dashed line) vs measurements (solid) for 1953~1998

tions agree in amplitude and phase with observations in the series except for 1990-94 while those for 3 seasons to come exhibit some difference and lag in both the aspects as compared to measurements.

2) SEASON-DEPENDENT VARIATION OF PREDICTIONS ETHODS

Seasonal curves were drawn, on a running month mean basis, for comparing the predictive skills (running means) of different leads (figure not shown). It is seen therefrom that higher (lowest) skill for NINO 1+2 is gained in predictions for mid summer to winter (spring) in advance. It is the lowest for predicting early summer (mid-summer to autumn) SST index in NINO3 (NINOs 4 and 3.4). On the whole, it seems that the lowest value exhibits westward propagation, a phenomenon that needs further study to see if it does bear a relation or not to the westward travel of genesis and development of an El Nino event.

IV. ASSESSMENT OF 1997/98 ENSO PREDICTIONS

The strongest El Nino episode in this century occurred in 1997/98. In the context of our model we made back forecasting and tracking predictions of the event, with the results given in Tab. 5.

It is seen from Tab. 5 that i) by the end of 1996, our model predicted the occurrence of the El Nino episode in mid-later 1997; ii) this episode was foreseen by prognoses made 1-4 seasons in advance before March 1997; iii) though the ENSO was predicted, the vigor was weaker. For instance, the prediction made in March 1997 for 3 seasons to come gave the SSTA of ≥1.2 compared to the measurement of >2.5; iv) predictions made before January 1998 for 2-4 seasons in advance indicate that the current El Nino episode would be likely to end after June 1998.

Tab. 5 Predictions, measured SST index and leading prognosis of the 1997/98 El Nino event for NINO3.4

Prediction starting from	Season(s) leading	Predicted period	Prediction anomaly	measured SST index
09-11/1996	4	09-11/1997	>+0.5	2.75
10-12/1996	4	10-12/1997	>+1.0	2.85
01-03/1997	1	04-06/1997	> +0.6	1.06
01-03/1997	2	07-09/1997	>+0.7	2.23
01-03/1997	3	10-12/1997	>+1.2	2.85
11/97-01/98	2	05-07/1998	<-0.2	
11/97-01/98	3	08-10/1998	<-1.0	
03-05/1998	2	09-11/1998	<-1.0	
03-05/1998	3	12/98-02/99	<-2.2	
06-08/1998	2	12/98-02/99	<-2.6	
06-08/1998	3	03-05/1999	<-1.8	

Besides, the 2-3 season leading predictions using data before May 1998 (Tab. 5) suggest that a stronger La Nina event would take place around the fall of 1998, and possibly reach its peak early in 1999. So far lots of evidence have shown that the 1997/98 El Nino event would end by June 1998, probably immediately followed by a La Nina episode.

V. DISCUSSION AND SUMMARY

a. The independent sample predictions, 1981-98 back forecasting, and the 1997-98 ENSO prediction all demonstrate high applicability of our model and scheme that are quite steady during operation. Take NINOs 3.4 and 4 predictions for example. Their 1-4 season leading predictions gave the correlations of greater than 0.85, 0.70, 0.50 and 0.35, respectively.

b. Our model compares advantageously to the CCA statistical model developed in the US CPC (Climate Prediction Center, Barnston et al., 1992) as regards predictive skill, and part of our predictions has exceeded those of the CPC model. In reference to 2-season leading predictions, for instance, the CCA model gives maximum scoring ranging over 0.85-0.89 just for wintertime months compared to the maximum of 0.85-0.87 for all seasons from our model, with the minimum of 0.30-0.35 (from the CCA) versus 0.37 (from our model).

Particularly, it should be pointed out that the CCA model requires a large volume of gridded data, consisting of global sea level pressure (SLP), tropical Pacific SST and sea level height (SLH), and the 20 ° C isothermal depth in contrast to 20 predictors for our model, leading to the much higher efficiency compared to the CCA model.

c. The paper presents a canonical mixed regression model which is actually the generalized simple type with scattered coefficients. The predictive scheme includes EEOF technique for the information on previous SST evolution (suggestive of the CCA establishment of a certain similarity relation between the previous SST evolution and the SST index at the prognostic time interval), PRESS sorting out model parameters and factors for independent sample experiments with 1-4 season leading predictions (suggesting a consensus prediction technique). All these have led to successful prediction of, say, the 1997/98 ENSO episode for longer than 6 months in advance and quite good prognosis of the termination of 1998 warm phase by the end of June based on measured data prior to January 1998. Preliminary prediction has been made of a strong La Nina event to occur in October 1998 whose precursors were beginning to emerge in the context of data before May 1998.

REFERENCES:

BARNSTON A G, ROPELEWSKI C F, 1992. Prediction of ENSO episodes using CCA [J]. *J.Climate*, **5**: 1316-1345.

DING Yu-guo, JIANG Zhi-hong, 1996. Study on canonical autoregression prediction of meteorlogical element fields [J]. *Acta. Meteor.Sin.*, **10**(1): 41-51

NOAA/NWS/NCEP, 1997. Climate Diagnostics Bulletin [R], 3.

PENLAND C T M, 1993. Prediction of Nino3 sea surface temperature using linear inverse modeling [J]. *J. climate*, **6**: 1067-1076.

XU J S, STORCH H, 1990. Principal oscillation patterns-prediction of state of ENSO [J] *J .Climate*, **3**: 1316-1429.

YAO Di-rong et al., 1992. A stepwise algorithm of selecting predictors following PRESS criterion (in Chinese) [J]. *Sci. Atmos. Sin.*, **16**(2):129-135.

ZEBIAK S E, CANE M A, 1987. A model El Nino-Southern Oscillation [J]. *Mon. Wea. Rev.*, **115**: 2262-2278.