

Application of linear dynamical mode decomposition to ensembles of climate simulations

Maria Buyanova¹, Sergey Kravtsov^{1,2}, Andrey Gavrilov¹, Dmitry Mukhin¹, Evgeny Loskutov¹
and Alexander Feigin¹

¹Institute of Applied Physics of the Russian Academy of Sciences, Nizhny Novgorod, Russian Federation

²University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA

Application of linear dynamical mode decomposition to ensembles of climate simulations

Maria Buyanova¹, Sergey Kravtsov^{1,2}, Andrey Gavrilov¹, Dmitry Mukhin¹, Evgeny Loskutov¹ and Alexander Feigin¹

¹Institute of Applied Physics of the Russian Academy of Sciences, Nizhny Novgorod, Russian Federation

²University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA

An analysis of the climate system is usually complicated by its very high dimensionality and its nonlinearity which impedes spatial and time scale separation. An even more difficult problem is to obtain separate estimates of the climate system's response to external forcing (e.g. anthropogenic emissions of greenhouse gases and aerosols) and the contribution of the climate system's internal variability into recent climate trends. Identification of spatiotemporal climatic patterns representing forced signals and internal variability in global climate models (GCMs) would make it possible to characterize these patterns in the observed data and to analyze dynamical relationships between these two types of climate variability.

In contrast with real climate observations, many GCMs are able to provide ensembles of many climate realizations under the same external forcing, with relatively independent initial conditions (e.g. LENS [1], MPI-GE [2], CMIP ensembles of 20th century climate). In this report, a recently developed method of empirical spatio-temporal data decomposition into linear dynamical modes (LDMs) [3,4] based on Bayesian approach, is modified to address the problem of self-consistent separation of the climate system internal variability modes and the forced response signals in such ensembles. The LDM method provides the time series of principal components and corresponding spatial patterns; in application to an ensemble of realizations, it determines both time series of the internal variability modes of current realization and the time series of forced response (defined as signal shared by all realizations). The advantage of LDMs is the ability to take into account the time scales of the system evolution better than some other linear techniques, e.g. traditional empirical orthogonal function decomposition. Furthermore, the modified ensemble LDM (E-LDM) method is designed to determine the optimal number of principal components and to distinguish their time scales for both internal variability modes and forced response signals.

The technique and results of applying LDM method to different GCM ensemble realizations will be presented and discussed. This research was supported by the Russian Science Foundation (Grant No. 18-12-00231).

[1] Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J., Bates, S., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J.-F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., and M. Vertenstein (2015), The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability, *Bulletin of the American Meteorological Society*, doi: 10.1175/BAMS-D-13-00255.1, 96, 1333-1349

[2] Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Dobrynin, M., Kornblueh, L., Kröger, J., Takano, Y., Ghosh, R., Hedemann, C., Li, C., Li, H., Manzini, E., Notz, N., Putrasahan, D., Boysen, L., Claussen, M., Ilyina, T., Olonscheck, D., Raddatz, T., Stevens, B. and Marotzke, J. (2019). The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability. *Journal of Advances in Modeling Earth Systems*, 11, 1-21. <https://doi.org/10.1029/2019MS001639>

[3] Gavrilov, A., Seleznev, A., Mukhin, D., Loskutov, E., Feigin, A., & Kurths, J. (2019). Linear dynamical modes as new variables for data-driven ENSO forecast. *Climate Dynamics*. 52(3–4), 2199–2216. <http://doi.org/10.1007/s00382-018-4255-7>

[4] Gavrilov, A., Kravtsov, S., Mukhin, D. (2020). Analysis of 20th century surface air temperature using linear dynamical modes. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(12), 123110. <https://doi.org/10.1063/5.0028246>

1. Data decomposition methods

- EOF decomposition

- LDM decomposition and optimization

- modified LDM method and its application to an ensemble of simulations (ELDM)

2. ELDM application to a model problem

- formulation and implementation

- results and comparison to other methods

- conclusions

- advanced problem formulation

3. ELDM application to LENS data (CESM Large Ensemble Community Project)

- data pre-processing

- implementation

- results

- discussion

4. Conclusions

Data decomposition methods

1. EOF (PCA) decomposition

Consider the spatiotemporal data $\mathbf{y}_n$ measured at N_T equidistant time moments with the spatial field $\mathbf{y}_n \in R^D$ arranged into column vectors; for gridded data, the dimension D is the number of spatial grid nodes. The EOF decomposition is the orthogonal rotation of $\mathbf{y}_n$,

$$\mathbf{y}_n = \mathbf{V} \mathbf{x}_n$$

such that the covariance matrix of the rotated time series $\mathbf{x}_n$ (called principal components, PCs), is diagonal. The columns of the orthogonal matrix $\mathbf{V}$ represent the spatial patterns (EOFs) associated with each PC; the PCs are arranged so that their variances are in descending order. The EOFs are eigenvectors of the covariance matrix of the original data with eigenvalues equal to the variances of the corresponding PCs $\mathbf{x}_n$.

2. Truncated EOF decomposition: for each d , reduction of data dimension from D to d ,

$$\mathbf{y}_n = \sum_{i=1}^d \mathbf{v}_i \cdot x_{in} + \eta_n$$

returns the minimum possible variance of the residual error η_n , which is the basis for using EOF decomposition for data compression required to lower computational time.

3. LDM decomposition [1, 2]:

Consider K leading PCs $\mathbf{x}_n$, obtained from data EOF decomposition:
$$\mathbf{x}_n = \mathbf{A} \cdot \mathbf{p}_n + \mathbf{c} + \sigma \boldsymbol{\xi}_n \quad (1)$$

where $\mathbf{p}$ is the time series of d new modes, and the columns of $K \times d$ matrix $\mathbf{A}$ are the new patterns; the corresponding LDM patterns in the physical space can be recovered from the matrix $\mathbf{V} \cdot \mathbf{A}$, where $\mathbf{V}$ is the truncated matrix of K leading EOFs, $\mathbf{c} \in R^K$, σ is scalar. The number d represents the LDM dimension. Each component of the residual $\boldsymbol{\xi}_n \in R^K$ is assumed to be an independent realization of the Gaussian white noise with zero mean and unit variance. All the parameters $\mathbf{A}$, $\mathbf{c}$, $\mathbf{p}$, and σ are unknown beforehand and are to be determined using the Bayesian framework.

LDM decomposition [1, 2]

The initial uncertainty of LDM parameters is described by their prior probability density functions (PDFs). The prior PDF for LDM mode time series $\mathbf{p}$ incorporates basic information about their time scales, which is at the heart of the method's capability to differentiate between slow and fast climate processes. P_N is the Gaussian PDF (3).

The vectors $\boldsymbol{\tau} \in R^d$ and $\boldsymbol{\sigma}_p \in R^d$ are hyperparameters, which assume that the i -th LDM mode p is a stochastic process with the autocorrelation time τ_i and variance σ_{pi}^2 . $N(0, 1)$ is a random number drawn from the standard normal distribution. The prior PDF for the LDM coefficients $\mathbf{A}, \mathbf{c}$ is taken to be Gaussian, ensuring, a priori, that the shape LDM covariance matrix matches that of the covariance matrix of the data $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_{N_T})$. The hyperparameters σ_A and σ_c are the scaling factors. The prior PDF for the parameter σ is a constant within some sufficiently wide interval and zero outside of this interval.

The important difference between LDM and EOF decompositions is in time scale determination. From the standpoint of LDM decomposition, the EOFs are generally sub-optimal: they correspond to the particular case when all the hyperparameters $\boldsymbol{\tau}$ are not optimized but instead set to zero (see [3]). This is due to the fact that while EOFs are designed to optimally capture the variance of the original data, the LDM procedure additionally considers the time scales of LDM mode time series to search for an optimal compromise between capturing the variance and spectral content of the data.

$$P_{pr}(\mathbf{P}|\boldsymbol{\tau}, \boldsymbol{\sigma}_p, d) = \prod_{i=1}^d \left[P_{\mathcal{N}}(p_{i,1}; \sigma_{pi}^2) \times \prod_{n=1}^{N-1} P_{\mathcal{N}}(p_{i,n+1} - \alpha_i \cdot p_{i,n}; \sigma_{pi}^2 \cdot (1 - \alpha_i^2)) \right] \quad (2)$$

$$\alpha_i \equiv \exp(-1/\tau_i).$$

$$P_{\mathcal{N}}(x; \sigma_x^2) := \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{x^2}{2\sigma_x^2}\right) \quad (3)$$

$$p_{i,n+1} = \alpha_i \cdot p_{i,n} + \sigma_{pi} \cdot \sqrt{1 - \alpha_i^2} \cdot \mathcal{N}(0, 1) \quad (4)$$

$$P_{pr}(\mathbf{A}, \mathbf{c}|\sigma_A, \sigma_c, d) = \prod_{k=1}^K \left[\prod_{i=1}^d P_{\mathcal{N}}(A_{ki}; \lambda_k \sigma_A^2) \cdot P_{\mathcal{N}}(c_k; \lambda_k \sigma_c^2) \right] \quad (5)$$

LDM method: optimization [1, 2]

Given all the prior PDFs, Bayesian formalism provides two cost functions for the search of the optimal LDM model. The first cost function is a posterior PDF for the LDM parameters $\boldsymbol{\mu} = (\mathbf{A}, \mathbf{c}, \mathbf{P}, \sigma)$ given the data $\mathbf{X}$, model (1), and the prior PDFs with the fixed hyperparameters $\mathbf{H} = (d, \boldsymbol{\tau}, \boldsymbol{\sigma}_p, \sigma_A, \sigma_c)$ (6). The hyperparameters $\mathbf{H}$ are also unknown beforehand. To optimize them, a second cost function is required — marginal likelihood (or Bayesian evidence) of the data given the hyperparameters $\mathbf{H}$ (7). The optimal hyperparameters $\mathbf{H}$ maximize (7) and provide the optimal values of LDM dimension d and time scales $\boldsymbol{\tau}$ whose likelihood is maximized given the data while avoiding the overfitting.

The parameters $\boldsymbol{\mu}$ maximizing (6) at the optimal (fixed) values of the hyperparameters $\mathbf{H}$ determine the optimal LDM model. The core PDF $(\boldsymbol{\mu}, \mathbf{X} | \mathbf{H})$ needed to estimate (6) and (7) can be found from (1) to (5) using (8-10).

To obtain the solution maximizing the functionals (6) and (7), a modified and improved version of the numerical algorithm originally developed in [3] is used.

$$P(\boldsymbol{\mu} | \mathbf{X}, \mathbf{H}) = \frac{P(\boldsymbol{\mu}, \mathbf{X} | \mathbf{H})}{P(\mathbf{X} | \mathbf{H})} \quad (6)$$

$$P(\mathbf{X} | \mathbf{H}) = \int P(\boldsymbol{\mu}, \mathbf{X} | \mathbf{H}) d\boldsymbol{\mu} \quad (7)$$

$$P(\boldsymbol{\mu}, \mathbf{X} | \mathbf{H}) = P(\mathbf{X} | \boldsymbol{\mu}, \mathbf{H}) \cdot P_{pr}(\boldsymbol{\mu} | \mathbf{H}), \quad (8)$$

$$P(\mathbf{X} | \boldsymbol{\mu}, \mathbf{H}) = \prod_{n=1}^N \prod_{k=1}^K P_{\mathcal{N}} \left(x_{k,n} - \sum_{i=1}^d A_{k,i} p_{i,n} - c_k; \sigma^2 \right) \quad (9)$$

$$P_{pr}(\boldsymbol{\mu} | \mathbf{H}) = P_{pr}(\mathbf{P} | \boldsymbol{\tau}, \boldsymbol{\sigma}_p, d) \times P_{pr}(\mathbf{A}, \mathbf{c} | \sigma_A, \sigma_c, d) \cdot P_{pr}(\sigma). \quad (10)$$

LDM modification: taking forcing response into account self-consistently

The original LDM method was adapted [2] to determine the system's response to the external forcing.

Consider the time series of the forced signal $\mathbf{f}$ given along with the data $\mathbf{x}$ ($\mathbf{f}$ can be derived, for example, from ensemble simulations of climate GCMs [2]). The forced signal can be treated as a prescribed mode with its own spatial pattern to enter the modified formulation of the LDM mode:

$$\mathbf{x}_n = \mathbf{A}\mathbf{p}_n + \mathbf{B}\mathbf{f}_n + \sigma\boldsymbol{\xi}_n + \mathbf{c}, \quad \mathbf{p}_n^{(i)} \in R^d, \quad \mathbf{f}_n \in R^{df}$$

The columns of the matrix $\mathbf{B}$ ($K \times df$ unknown parameters) define spatial patterns corresponding to different forcing response modes $\mathbf{f}$.

The same Bayesian formalism is applied to find the optimal values of the parameters $\boldsymbol{\mu} = (\mathbf{A}, \mathbf{B}, \mathbf{c}, \mathbf{P}, \sigma)$ in the extended LDM method. The prior PDFs for $\mathbf{A}$, $\mathbf{c}$, $\mathbf{P}$, σ are the same, and the prior PDF for $\mathbf{B}$ (11) is in full analogy with (5).

$$P_{pr}(\mathbf{B}|\sigma_B) = \prod_{k=1}^K \prod_{i=1}^{df} P_{\mathcal{N}}(B_{ki}; \lambda_k \sigma_B^2). \quad (11)$$

$$P(\mathbf{X}|\boldsymbol{\mu}, \mathbf{H}) = \prod_{n=1}^N \prod_{k=1}^K P_{\mathcal{N}} \left(x_{kn} - \sum_{i=1}^d A_{ki} p_{in} - \sum_{i=1}^{df} B_{ki} f_{in} - c_k; \sigma^2 \right) \quad (12)$$

In this procedure, forcing response patterns $\mathbf{B}$ automatically are internally consistent with the other parameters of the LDM decomposition, which is an improvement over the traditional strategy of removing the forced signal using linear regression prior to any data analysis.

$$P_{pr}(\boldsymbol{\mu}|\mathbf{H}) = P_{pr}(\mathbf{P}|\boldsymbol{\tau}, \sigma_p, d) \cdot P_{pr}(\mathbf{A}, \mathbf{c}|\sigma_A, \sigma_c, d) \cdot P_{pr}(\mathbf{B}|\sigma_B) \cdot P_{pr}(\sigma). \quad (13)$$

ELDM application to an ensemble of data simulations

It is assumed that

- The external forcing presents in all realizations (simulation runs) and each time affects the system in a similar way. The system's forcing response is then determined as a part of the data that appears in all realizations. Both forcing response time series f and spatial patterns B are assumed shared for all realizations.
- Internal variability modes represent a part of the data which is unique for each realization. Their time series are specific, but the corresponding spatial patterns are assumed shared for all realizations. (For example, this may represent a process affecting the same grid nodes but starting from different initial conditions for many simulation runs.)

Data pre-processing:

1. Consider the data $\mathbf{y}_n \in R^D$ defined at N_T equidistant time moments and arranged into D column vectors; for gridded data, the dimension D is the number of spatial grid nodes. N_R realizations of $\mathbf{y}_n$ (simulation runs) are assumed independent.
2. The data $\mathbf{y}_n$ of all realizations are rearranged in long column vectors by stacking time series of all realizations consecutively one by one so that complete data dimensions are $[N_R \times N_T, D]$. Thus all realizations are processed together for each of D spatial grid nodes. The long columns of data are centered and normalized to have zero mean and unit variance for each grid node.

$$\mathbf{y}_n^{(i)} \rightarrow \mathbf{Y}_m \ (m = 1, \dots, N_T \times N_R)$$

1. The EOF decomposition is applied to the rearranged data, and data compression is performed then, leaving $D' \ll D$ leading PCs for further processing: $\mathbf{Y}_m = \mathbf{V} \mathbf{X}_m + \boldsymbol{\zeta}_m$, $\mathbf{X}_m \in R^{D'}$.
2. The PCs data can be then rearranged back as $[N_R, N_T, D']$ -dimensions array in order to separate single realizations: $\mathbf{X}_m \rightarrow \mathbf{x}_n^{(i)}$. Thus each i -th component of $\mathbf{x}_n^{(i)}$ reflects the specifics of each realization, however the spatial patterns described by corresponding EOFs $\mathbf{V}$ remain shared for all realizations.
3. The ELDM method is then applied to the complete set of $\mathbf{x}_n^{(i)}$. Time series of the resulting modes, $\mathbf{p}_n^{(i)}$, are specific for each realization; forcing response time series are shared. Spatial patterns of the resulting modes in $\mathbf{y}$ data space are $\mathbf{A}' = \mathbf{V} \mathbf{A}$, $\mathbf{B}' = \mathbf{V} \mathbf{B}$.

Model problem: formulation and implementation

To prove the ELDM method restores the actual characteristics of the data, a model problem was designed and solved:
given initial modes $\rightarrow$ model ensemble data $\mathbf{y}_n^{(i)} \rightarrow$ modes restored by ELDM

The given modes are chosen so to produce a realistic ensemble data (imitating real or simulated climate data obtained from GCMs):

- several modes having different time scales should be taken into account;
- ratios between the terms in data $\mathbf{y}$ related to different data components (slow p , fast p , forcing response f , noise) should be similar to those of realistic data
- 10 realizations are processed; it should be checked whether this number of realizations can provide a reliable result for realistic data of similar characteristics

A model problem with an additive forcing response was tested: $\mathbf{p}_n^{(i)}, f_n \rightarrow \mathbf{y}_n^{(i)} \rightarrow \tilde{\mathbf{p}}_n^{(i)}, \tilde{f}_n$.

Consider a system which demonstrates 2 internal modes p (“slow” and “fast”) and a external forcing response signal f ; all these signals are modelled using a discrete stochastic oscillator:

$$(*) \quad \begin{pmatrix} u_n \\ v_n \end{pmatrix} = e^{-\gamma} \begin{pmatrix} \cos \omega & \sin \omega \\ -\sin \omega & \cos \omega \end{pmatrix} \begin{pmatrix} u_{n-1} \\ v_{n-1} \end{pmatrix} + \begin{pmatrix} \sigma_u \xi_n^u \\ \sigma_v \xi_n^v \end{pmatrix}, \text{ where } \gamma = 1/\tau, \omega = 2\pi\gamma,$$

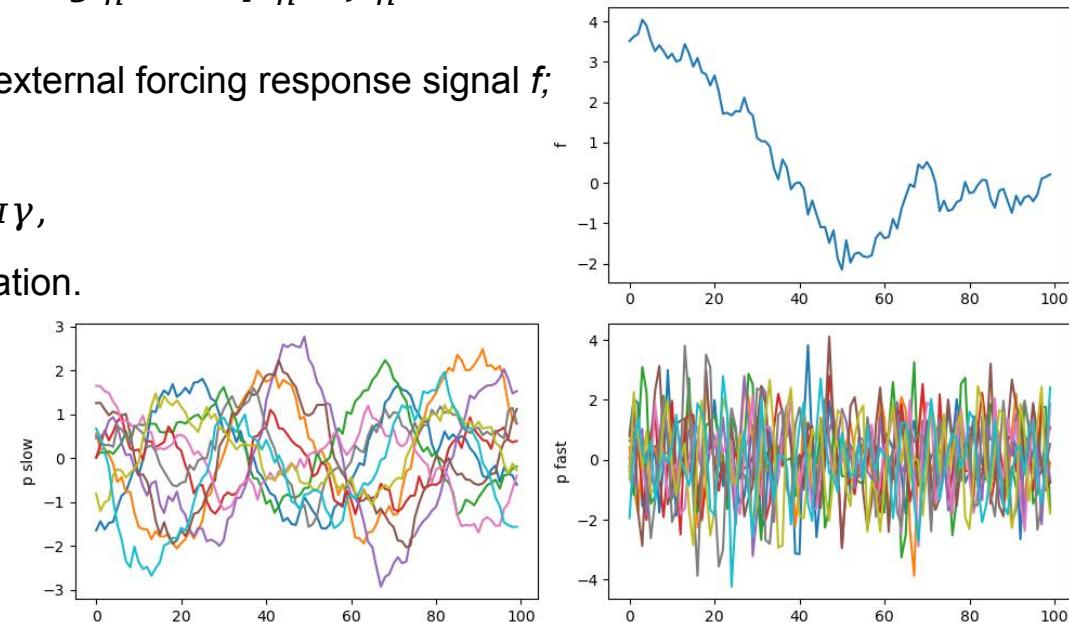
τ describes the time scale, ξ_n^u, ξ_n^v are white noise signals, unique for each realization.

$$\tau_{p \text{ slow}} = 50, \sigma_{(u, v)p \text{ slow}} = 0.2 \rightarrow u \text{ gives } p^{\text{slow}}$$

$$\tau_{p \text{ fast}} = 5, \sigma_{(u, v)p \text{ fast}} = 0.6 \rightarrow u \text{ gives } p^{\text{fast}}$$

$$\tau_f = 100, \sigma_{(u, v)f} = 0.2 \rightarrow u \text{ gives } f$$

10 realizations of internal modes are modelled.



Model problem: formulation and implementation

The ensemble model data are defined as $\mathbf{y}_n^{(i)} = \mathbf{A}\mathbf{p}_n^{(i)} + \mathbf{B}\mathbf{f}_n + \sigma\boldsymbol{\xi}_n^{(i)} + \mathbf{c}$. $n = 1, \dots, N_T$, $i = 1, \dots, N_R$

Spatial patterns $\mathbf{A}$ and $\mathbf{B}$ are modelled by sets of random values for $D=20$ grid nodes.

Random white noise (normal distribution with zero mean and unit variance) $\boldsymbol{\xi}_n^{(i)}$ is added, $\mathbf{c} \in R^D$.

The data are arranged as long columns (by stacking all realizations consecutively) and the ensemble EOF decomposition is then applied:

$$\mathbf{y}_n^{(i)} \rightarrow \mathbf{Y}_m \ (m = 1, \dots, N_T \times N_R); \quad \mathbf{Y}_m = \mathbf{V}\mathbf{X}_m, \quad \mathbf{X}_m \in R^D; \quad \mathbf{X}_m \rightarrow \mathbf{x}_n^{(i)}$$

Unlike the case of real data processing, no data compression is applied: all ensemble PCs are taken as input parameters for ELDM.

Grid nodes number D is used as the number of processed ensemble PCs.

Additional condition is applied to the ratios of the $\mathbf{y}$ terms variances (to correspond to real data term ratios, e.g. for the near-surface air temperature data used in [2]), calculated for data taken as long columns:

$$\text{var}(\mathbf{A}\mathbf{p}_n^{slow}) : \text{var}(\mathbf{A}\mathbf{p}_n^{fast}) : \text{var}(\mathbf{B}\mathbf{f}_n) : \text{var}(\sigma\boldsymbol{\xi}_n) = 0.2 : 0.25 : 0.25 : 0.3$$

The ELDM method is then applied to the complete set of $\mathbf{x}_n^{(i)}$ to restore the internal variability modes $\tilde{\mathbf{p}}_n^{(i)}$ and forcing response modes $\tilde{\mathbf{f}}_n$:

$$\mathbf{x}_n^{(i)} = \tilde{\mathbf{A}}'\tilde{\mathbf{p}}_n^{(i)} + \tilde{\mathbf{B}}'\tilde{\mathbf{f}}_n + \sigma\tilde{\boldsymbol{\xi}}_n^{(i)}, \quad n = 1, \dots, N_T, \quad i = 1, \dots, N_R; \quad \tilde{\mathbf{p}}_n^{(i)} \in R^d, \quad \tilde{\mathbf{f}}_n \in R^{df}$$

Time series of the resulting modes in $\mathbf{y}$ data space are $\tilde{\mathbf{p}}_n^{(i)}$ and $\tilde{\mathbf{f}}_n$. Spatial patterns of the resulting modes are $\tilde{\mathbf{A}} = \mathbf{V}\tilde{\mathbf{A}}'$, $\tilde{\mathbf{B}} = \mathbf{V}\tilde{\mathbf{B}}'$.

Several combination of assumed forcing response dimension df and internal variability modes dimension d are studied. The optimal parameters (providing maximization of the modified LDM method's cost functions, with analogy to (6) and (7)) and resulting dimensions are to be compared to the initial given data.

Model problem: other methods to separate forcing response and determine the internal variability modes

Two simple methods based on EOF decomposition were also tested:

1. Extracting the ensemble mean

Independent realisations of internal variability in different runs tend to cancel in the ensemble mean, therefore the ensemble mean may represent forced term $\mathbf{y}_f$ in data (separately for each grid node). After extracting this term, the ensemble EOF decomposition can be used to describe the main characteristics of the internal variability: PCs act as the internal modes time series (specific for each realization) and corresponding EOFs represent their spatial patterns (shared for all realizations). Limitation of this method is the inability to separate the time series and the patterns of the forcing response.

2. Using “spatial” PCs

It is assumed the external forcing affects the system in all D grid nodes and in all realizations. The EOF decomposition is applied to the data arranged as $[N_T, N_R \times D]$ (thus merging realizations and grid nodes to the new “spatial” dimension). The obtained “spatial” PCs are considered as forcing response time series. The number of required PCs (the forcing response modes dimension) is unknown; several values should be tested. After extracting their total summary term $\mathbf{y}_f$ from the data, the ensemble EOF decomposition can be applied to the residual data in the same way as described above in order to obtain internal variability modes and their spatial patterns.

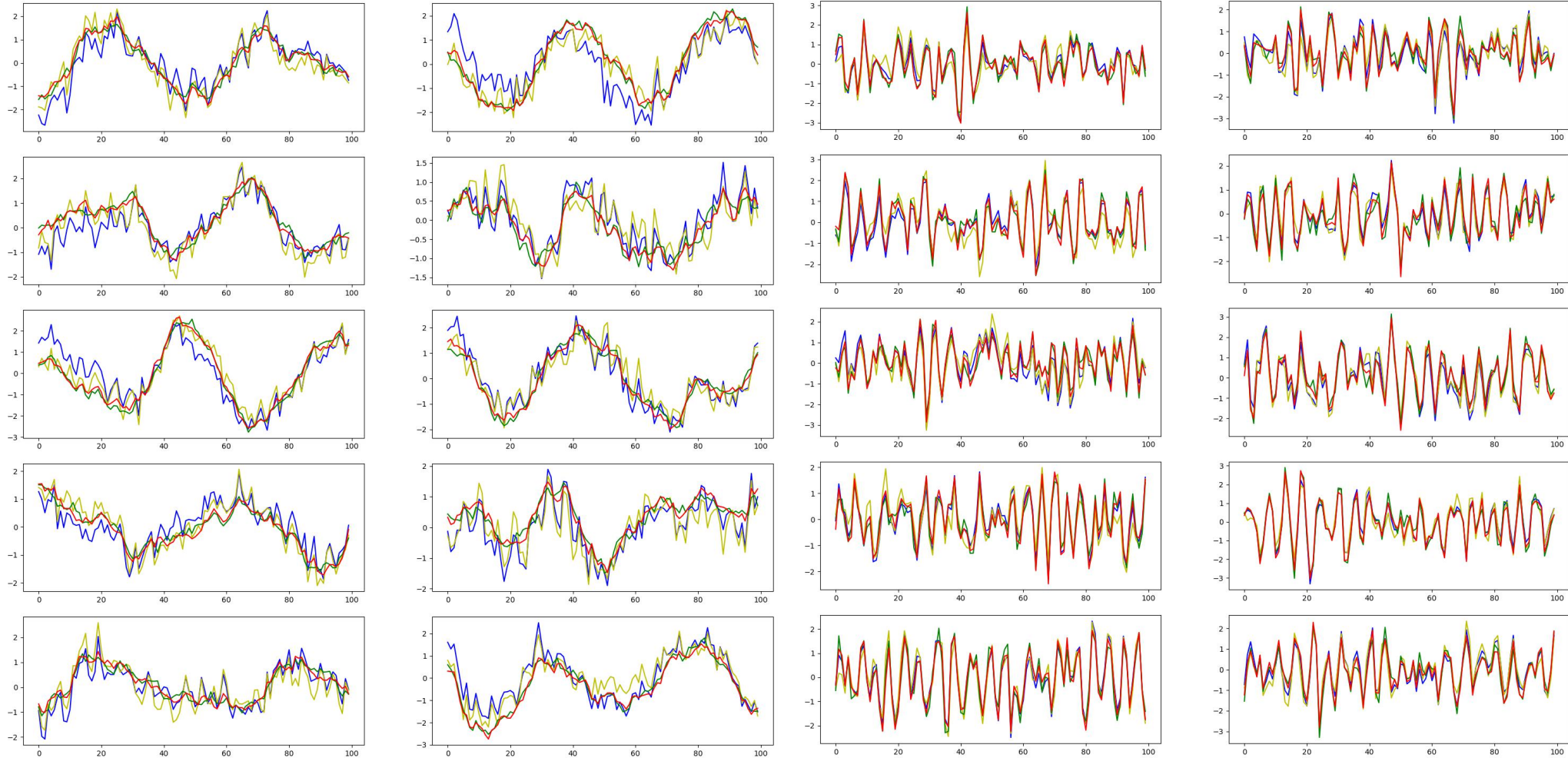
Both methods were applied to the model problem to compare their results to those of the ELDM method and to the initial data. The ELDM optimized dimension of the internal variability modes was found as $d=2$, and the forcing response dimension $df=1$; both coincide to initial given values. It was also obtained that the spatial PCs method also gives most appropriate result for 1 PC taken as the forcing response mode.

Model problem: results comparison

Internal modes time series p for 10 realizations, normalized by standard deviation (calculated through all realizations):

- initial (green),
- restored by the ELDM method (red),
- computed as ensemble PCs after extracting the ensemble mean taken as the forced term y_f (yellow),
- computed as ensemble PCs after extracting the terms related to m leading spatial PCs taken as forcing response modes (best for $m=1$: providing the most correct internal modes determination, blue)

left: slow p mode
right: fast p mode

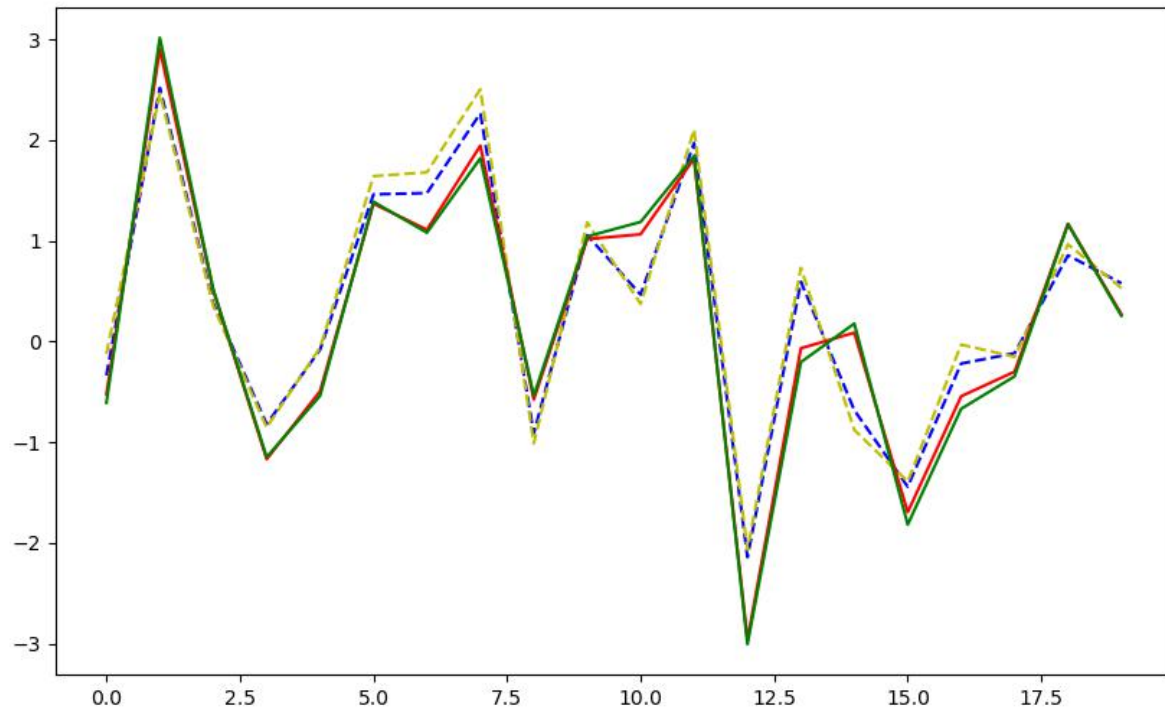


Model problem: results comparison

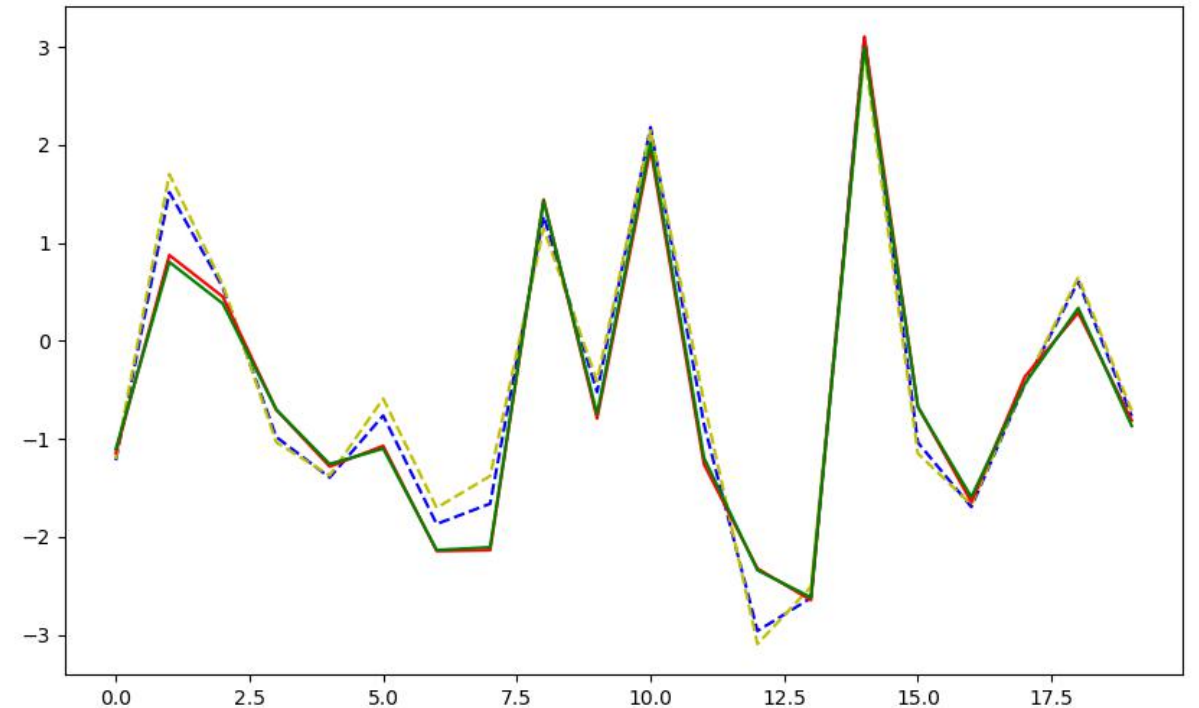
Internal modes spatial patterns A (20 values in data space, corresponding to 20 grid nodes, arranged in one line):

- initial (green),
- restored by ELDM method (red),
- computed as ensemble EOFs after extracting the ensemble mean as the forced term $\mathbf{y}_f$ (yellow),
- computed as ensemble EOFs after extracting the terms related to m leading spatial PCs taken as forcing response modes (blue, best* for $m=1$)

p -mode patterns A , multiplied by the corresponding mode standard deviation (calculated through all realizations):
Slow mode



Fast mode



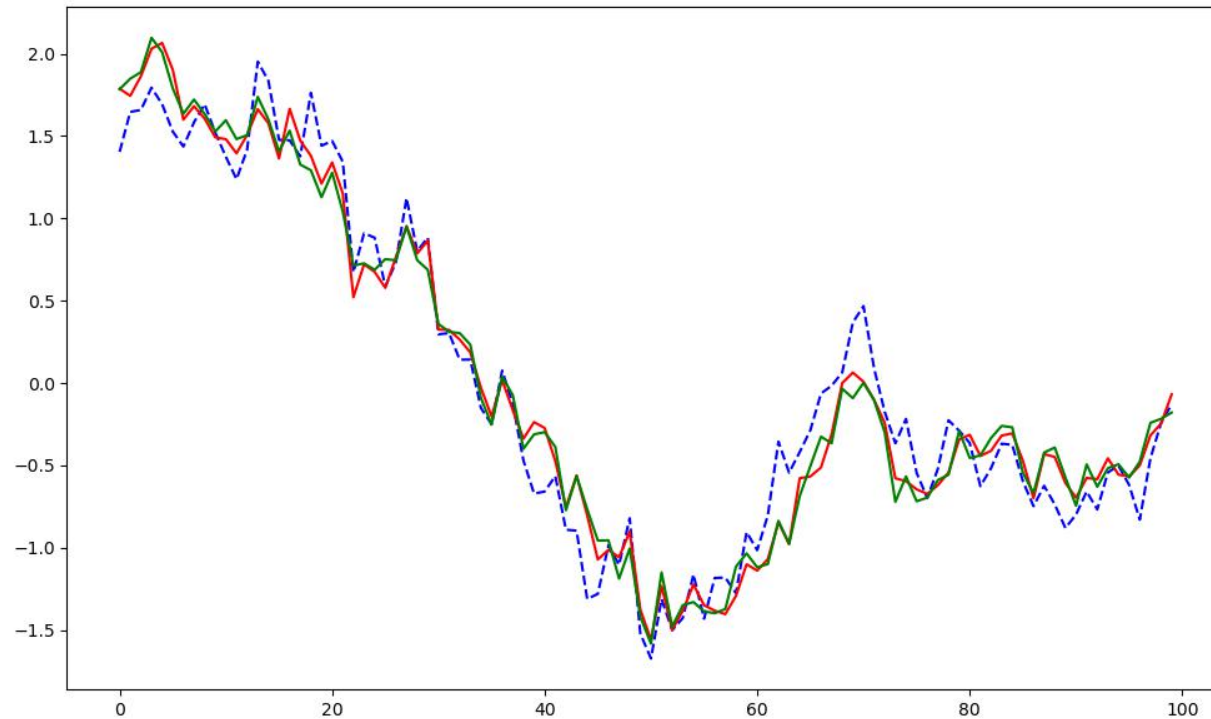
*the value providing the most correct internal modes determination

Model problem: results comparison

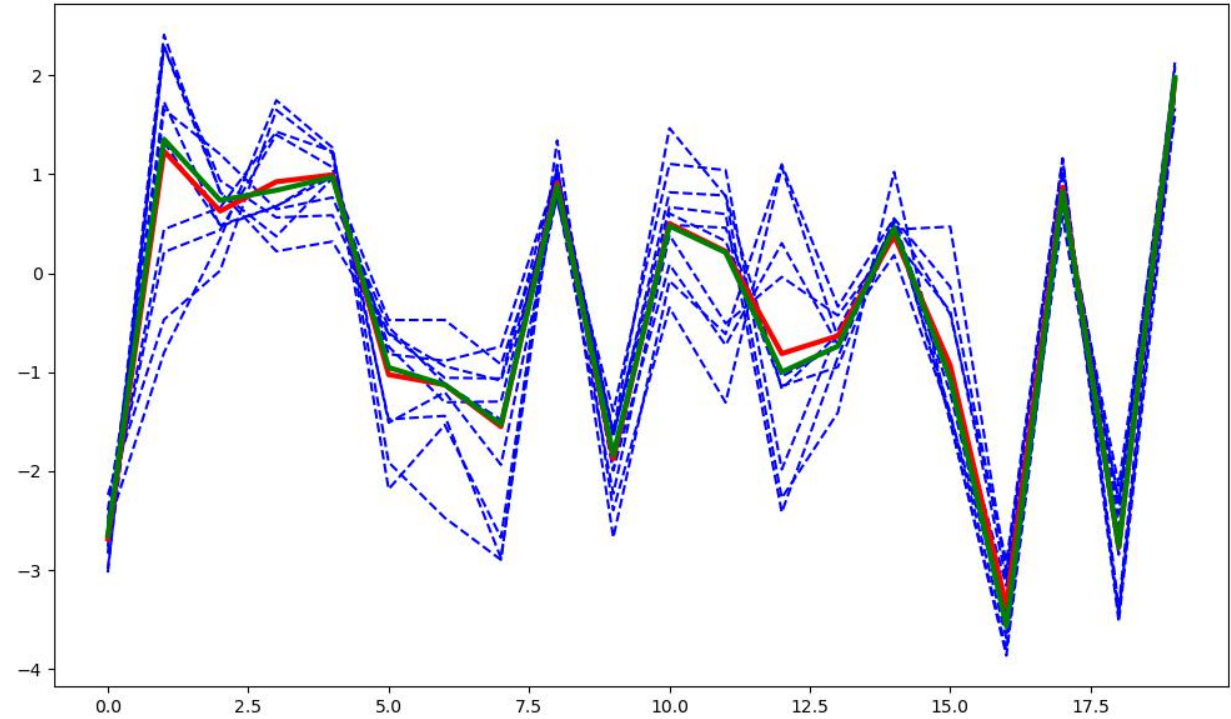
Forcing response time series f and spatial patterns B (20 values in data space, corresponding to 20 grid nodes, arranged in one line):

- initial (green),
- restored by ELDM method (red),
- m leading spatial PCs taken as forcing response modes and their patterns (blue, best* for $m=1$): according to the spatial PCs method, each realization has its unique spatial pattern corresponding to the same spatial PC mode taken as f .

Forcing response time series f (normalized by standard deviation)



Patterns B (multiplied by f standard deviation):



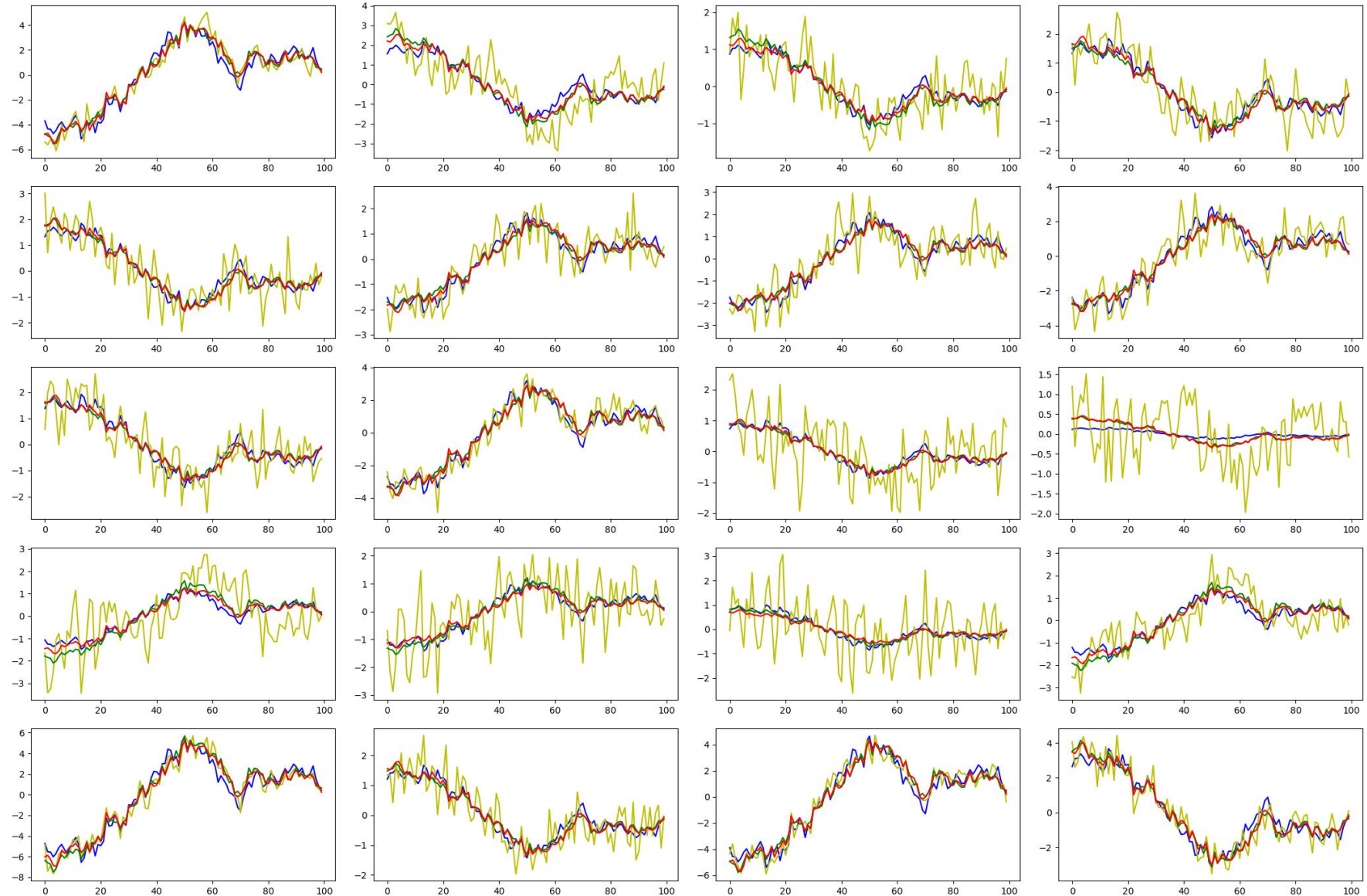
*the value providing the most correct internal modes determination

Model problem: results comparison

Forced term in data, $\mathbf{y}_{f\ n} = \mathbf{B}\mathbf{f}_n$, for 20 spatial components (20 grid nodes) separately, in time:

- initial (green),
- restored by ELDM (red),
- ensemble mean (yellow),
- sum of the terms related to m leading spatial PCs taken as forcing response modes (blue, best* result for $m=1$)

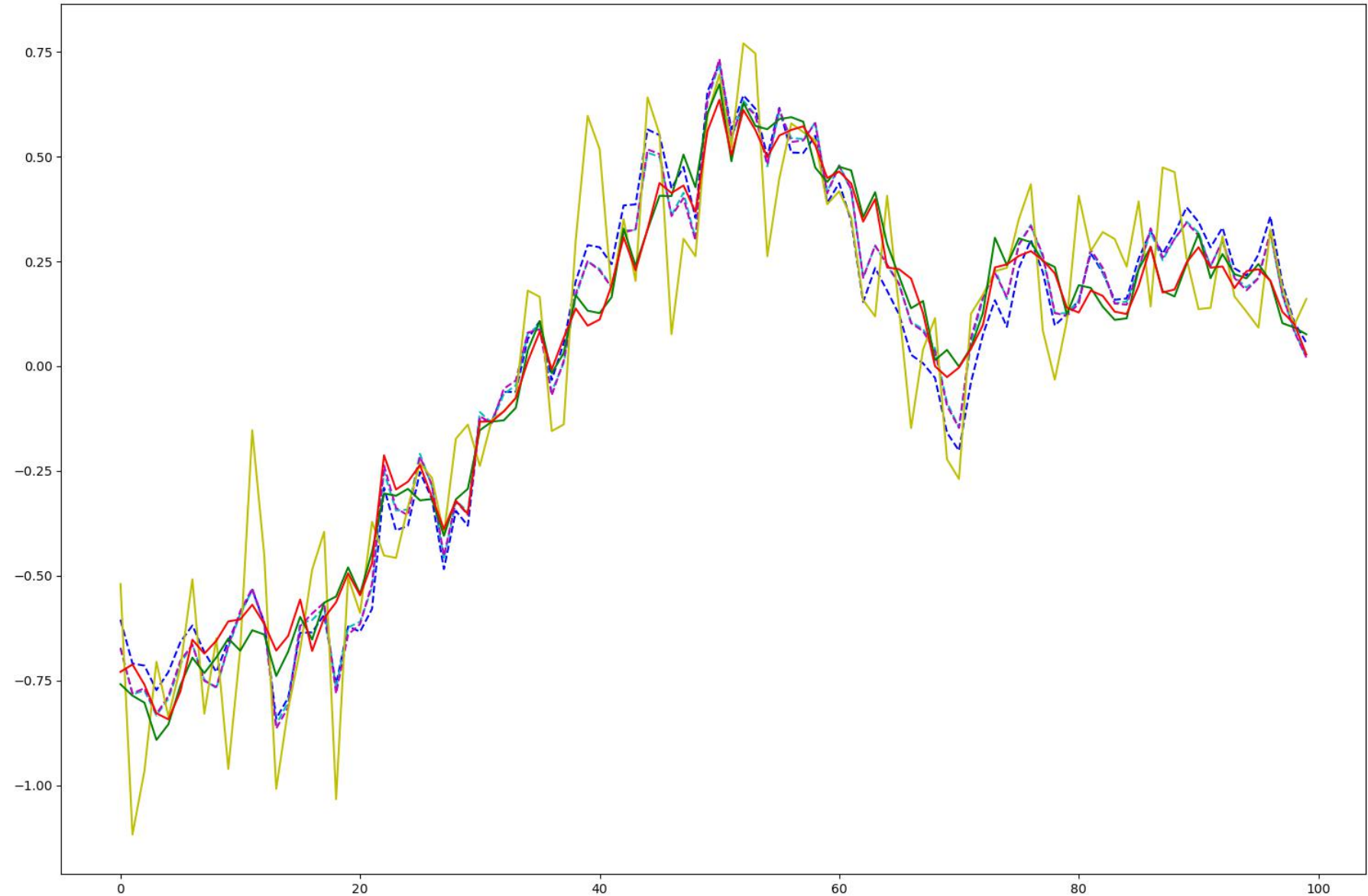
*the value providing the most correct internal modes determination



Model problem: results comparison

Forced term in data, $\mathbf{y}_{f\ n} = \mathbf{B}\mathbf{f}_n$, averaged over 20 spatial components (20 grid nodes), in time:

- initial (green),
- restored by ELDM method (red),
- ensemble mean (yellow),
- sum of the terms related to m leading spatial PCs taken as forcing response modes
($m=1$, blue)
($m=2$, cyan)
($m=3$, magenta)



Model problem: results comparison and conclusions

Energetical characteristics of different terms in y corresponding to various internal variability modes, the forcing response and the noise component: ELDM-restored terms variances agrees with the initial values (both calculated through all realizations).

	Slow p component: ($A p$)	Fast p component ($A p$)	Summary forced term:	Noise term (as data minus summary forced and internal variability terms):
Initial: term variance / total y variance	0.20	0.25	0.25	0.30
ELDM result: term variance / total y variance	0.19	0.26	0.24	0.27

1. ELDM properly determines the forced term in data. The results are in good agreement with the initial given data and with results obtained from two other described methods of forcing response determination. In contrast to other methods, the ELDM method optimizes the forcing response dimension self-consistently. For such implementation of the forcing response, its explicit time series and spatial patterns can be determined; they coincide with the given data.
2. Unlike other described methods, ELDM optimizes also the internal variability modes dimension and restores correctly their time series together with their spatial patterns. The ELDM-restored internal variability modes demonstrate clear separation of the time scales, in contrast to results of two other methods based on EOF data decomposition.
3. ELDM method correctly restores the energetical characteristics of different data terms.

Model problem: advanced formulation

An advanced model problem was also studied, taking into account a multimodal forcing response given indirectly.

A similar system with 2 internal unforced modes (“slow” and “fast”) and several components of the forcing response is considered. One of the forcing response components is defined by the same additive mode f with corresponding spatial pattern $\mathbf{B}$, the rest represent the system's response to f signal. A discrete stochastic oscillator is used to model the complete $\mathbf{p}$ modes (which are accompanied by the same spatial patterns $\mathbf{A}$):

$$(**) \quad \begin{pmatrix} u_n \\ v_n \end{pmatrix} = e^{-\gamma} \begin{pmatrix} \cos \omega & \sin \omega \\ -\sin \omega & \cos \omega \end{pmatrix} \begin{pmatrix} u_{n-1} \\ v_{n-1} \end{pmatrix} + \begin{pmatrix} \sigma_u \xi_n^u \\ \sigma_v \xi_n^v \end{pmatrix} + \varepsilon f_n$$

$$\tau_{p \text{ slow}} = 50, \sigma_{(u, v) p \text{ slow}} = 0.2, \varepsilon_{\text{slow}} = 0.1 \rightarrow u \text{ gives } p^{\text{slow}}$$

$$\tau_{p \text{ fast}} = 5, \sigma_{(u, v) p \text{ slow}} = 0.6, \varepsilon_{\text{fast}} = 1 \rightarrow u \text{ gives } p^{\text{fast}}$$

Such formulation of the initial forcing response provides more realistic self-consistent modelling of $\mathbf{y}$ data, though it complicates comparison to ELDM results (a mode-to-mode forcing response matching is impossible because of “mixed” spatial patterns icorresponding to initial forced modes).

The forced term in initial data, $\mathbf{y}_f$, and the corresponding term in the ELDM resulting data were compared instead of separate forcing response time series and patterns. Their good agreement was shown; moreover, the ELDM results still were closer to the inital data then the results of two additional methods described here. ELDM method advantages were even more remarkable in determination of the internal variability modes. The details will be shown in the upcoming paper.

ELDM application to LENS (CESM LARGE ENSEMBLE COMMUNITY PROJECT [4, 5])

The Large Ensemble Project includes a 40-member ensemble of fully-coupled CESM1 simulations for the period 1920-2100. Each member is subject to the same radiative forcing scenario (historical up to 2005) and begins from a slightly different initial atmospheric state (created by randomly perturbing temperatures at the level of round-off error).

The ELDM method is applied to the near-surface air temperature data (TREFHT variable, $D = 190 \times 288$ spatial grid nodes) on the interval from 1920 till 2005. The main purpose is to study large time scale modes of internal variability, like the Global Stadium Wave [2], therefore the timestep of 1 year is chosen. 40 of the 41 independent simulation runs are taken into account, shuffled and divided into 4 groups of $N_R=10$ realizations (groups A, B, C, D). Each group is processed separately.

The time series of $N_T = 86$ values are studied: $\mathbf{y}_n^{(i)}, \mathbf{y}_n^{(i)} \in R^D$ ($n = 1, \dots, N_T, i = 1, \dots, N_R$).

In each group of realizations the data are rearranged as D long columns, $\mathbf{Y}_m \in R^D$ ($m = 1, \dots, N_T \times N_R$):

$$\mathbf{Y}_m = [\mathbf{y}_1^{(1)}; \mathbf{y}_2^{(1)}; \dots; \mathbf{y}_{N_T}^{(1)}; \mathbf{y}_1^{(2)}; \dots; \mathbf{y}_{N_T}^{(2)}; \dots; \mathbf{y}_{N_T}^{(N_R)}].$$

Each column corresponds to a single grid node; its data is centered to have zero mean, normalized by its standard deviation and multiplied by the square root of the latitude cosine.

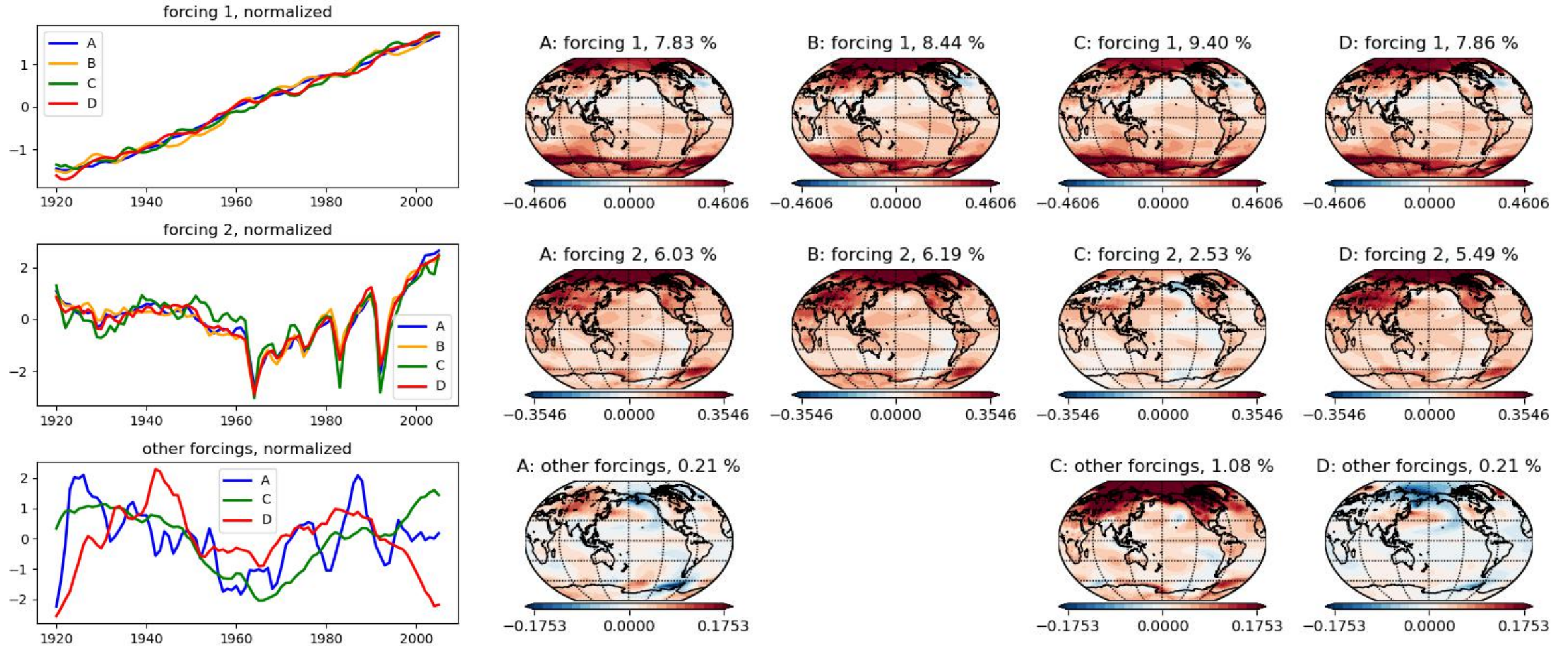
The ensemble EOF decomposition is applied to the rearranged data: $\mathbf{Y}_m = \mathbf{V}^T \mathbf{X}_m + \eta_m$, $\mathbf{X}_m \in R^K$, $K \ll D$; η_m is residual error.

$K=20$ leading PCs (complete set of $\mathbf{x}_n^{(i)} \in R^K$) are taken into account to be processed by the ELDM method.

Several combinations of assumed forcing response dimension df and internal variability modes dimension d are studied. The optimal parameters (providing maximization of the modified LDM method's cost functions, with analogy to (6) and (7)) and resulting dimensions obtained for a current group of realizations are compared to results of other groups.

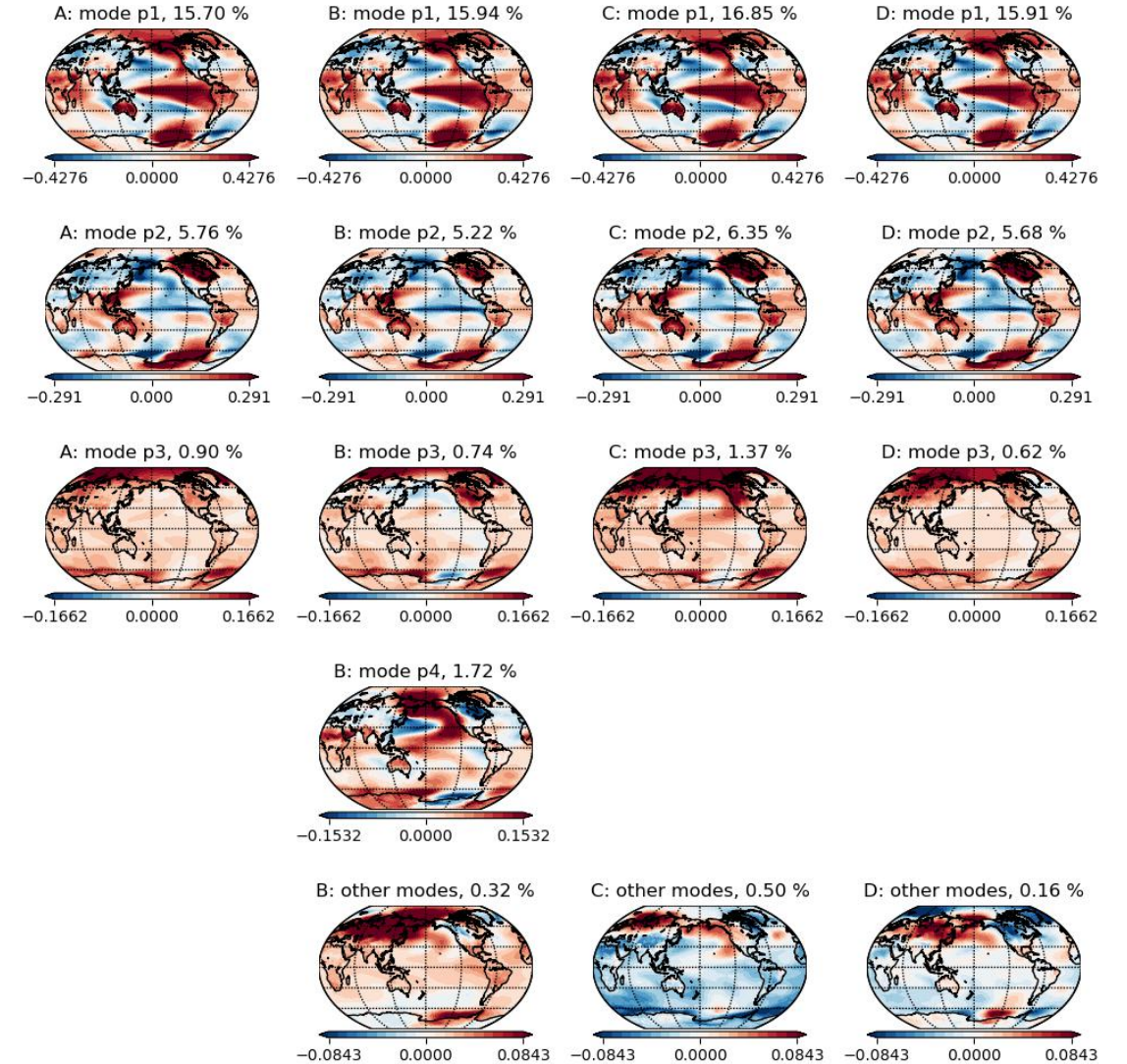
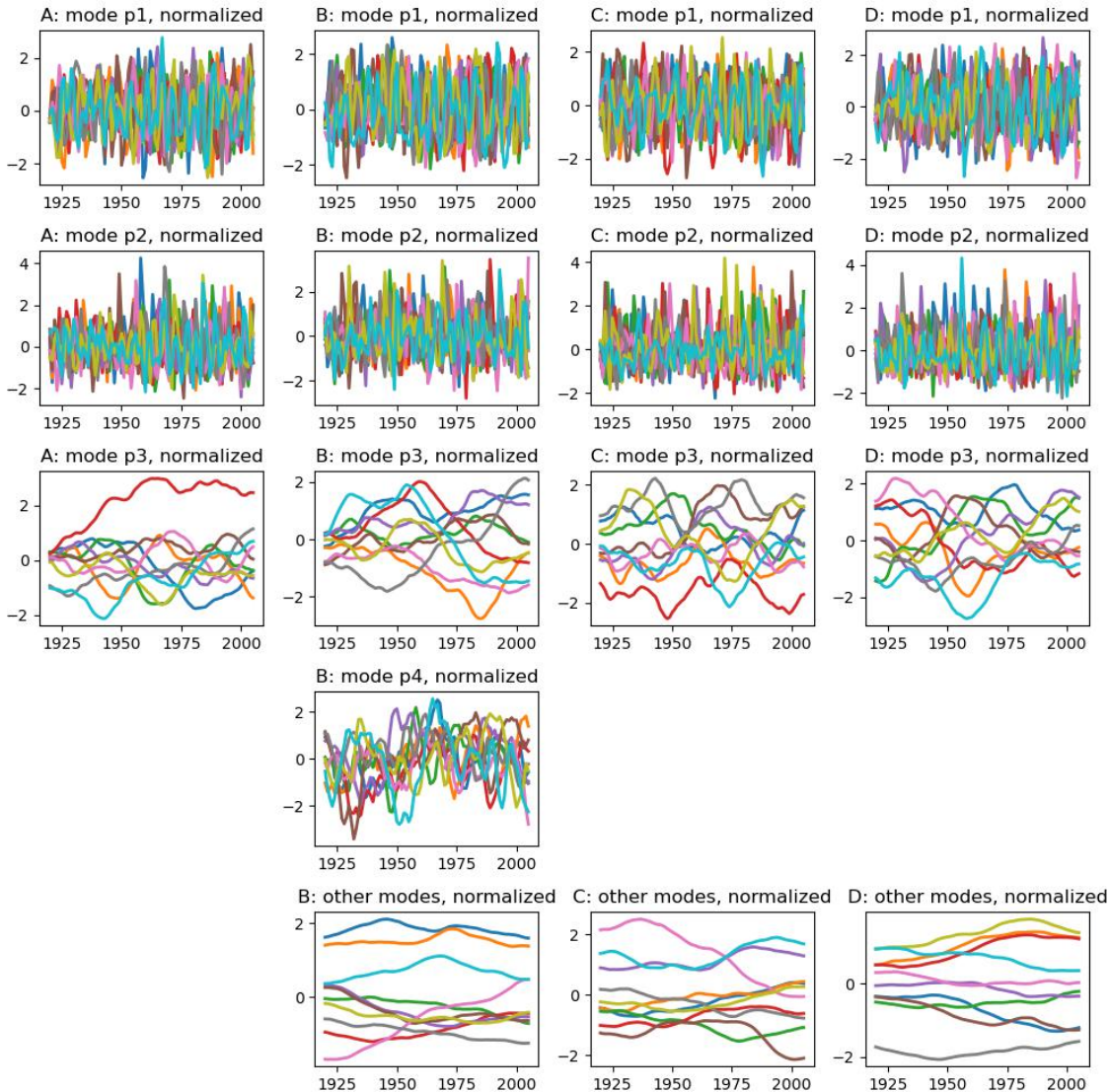
ELDM application to LENS: results

All 4 groups of 10 realizations (A, B, C, D) show some similar forcing response time series f as well as their spatial patterns (in degrees), though resulting forcing response dimension may differ. The variance of the term $\mathbf{B}f$ with respect to the total $\mathbf{y}$ data variance is specified at each pattern plot.



ELDM application to LENS : results

All 4 groups of realizations (A, B, C, D) show similar internal variability spatial patterns shared for all realizations in a group. The modes time series p have similar time scales (either within each group and between all groups), though each realization is unique. The variance of the term A_p with respect to the total y data variance is specified at each pattern plot.



ELDM application to LENS: discussion

- The most pronounced forcing response modes with the highest variances (the first and the second ones) found by the ELDM method coincide well in all groups. They represent a combination of the response to greenhouse gases and anthropogenic/volcanic aerosols.
- The spatial patterns of most pronounced modes of internal variability coincide well in all groups. The timescales of the modes are well separated. The first and the second modes of internal variability in each group are related to the El Nino - Southern Oscillation. The third mode in each group shows a high latitudes multidecadal process. Striking similarity between its spatial pattern and the leading forcing response modes' patterns was found, suggesting that the forced response strongly projects onto the low-frequency internal variability modes. Similar projection of a forcing response onto the low-frequency system's internal variability was studied in the advanced model problem (described earlier), when the ELDM-restored forcing response pattern also resembled the internal variability modes' patterns. The details will be shown in the upcoming paper.
- For major modes of forcing response and internal variability 10 is a sufficient number of realizations in an ensemble to obtain reliable results.

Disagreement in groups results: possible reasons

- For minor modes of forcing response and internal variability a higher number of realizations in an ensemble (>10) may be required. More detailed computations should be performed to get the estimate.
- Time line (86 year values) is not long enough.
- Only 67% of the total data y_n variance is captured by the leading 20 PCs used in computations: energy loss during data pre-processing (ensemble EOF decomposition) is noticeable. (This can be justified by the data specifics: the PC variance decreases very slow with the PC number, however each of the omitted PCs captures variance which is less than 1% of the total data variance.) Nevertheless, processing a much extended PCs set is computationally expensive.

Conclusions

The LDM method was adapted [2] to determine the system's response to the external forcing. Due to this modification, the forcing response is internally consistent with the other parameters of the LDM decomposition, which is an improvement over the traditional strategy of removing the forced signal using linear regression prior to any data analysis. The modified ensemble LDM method determines both time series and spatial patterns of the internal variability and of the forcing response as well. For both components of the system dynamics, ELDM inherits the ability of the original LDM method to take into account the time scales of the system evolution.

The model problem was formulated and solved. It was confirmed the ELDM method determines both internal variability and forcing response modes correctly. The advanced model problem was also studied to take into account the forcing response projection onto the system's internal variability and to illustrate the ELDM application to realistic data. Similar results were obtained for the LENS ensemble data, in 4 separate groups of realizations. The details will be shown in the upcoming paper.

References

- [1] Gavrilov, A., Seleznev, A., Mukhin, D., Loskutov, E., Feigin, A., & Kurths, J. (2019). Linear dynamical modes as new variables for data-driven ENSO forecast. *Climate Dynamics*. 52(3–4), 2199–2216. <http://doi.org/10.1007/s00382-018-4255-7>
- [2] Gavrilov, A., Kravtsov, S., Mukhin, D. (2020). Analysis of 20th century surface air temperature using linear dynamical modes. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(12), 123110. <https://doi.org/10.1063/5.0028246>
- [3] Gavrilov, A., Mukhin, D., Loskutov, E., Volodin, E., Feigin, A., and Kurths, J., “Method for reconstructing nonlinear modes with adaptive structure from multidimensional data,” *Chaos* 26, 123101 (2016).
- [4] Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J., Bates, S., Danabasoglu, G., Edwards, J., Holland, M. Kushner, P., Lamarque, J.-F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., and M. Vertenstein (2015), The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability, *Bulletin of the American Meteorological Society*, doi: 10.1175/BAMS-D-13-00255.1, 96, 1333-1349
- [5] <https://www.cesm.ucar.edu/projects/community-projects/LENS>