Review of
”Spy4Cast v1.0: a Python Tool for statistical seasonal forecast based on Maximum Covariance Analysis”

Pablo Duran-Fonseca and Belén Rodríguez-Fonseca

January 29, 2025

General Comments

This manuscript presents *Spy4Cast*, a Python-based library aimed at facilitating Maximum Covariance Analysis (MCA) in climate research, especially for seasonal prediction and teleconnection studies. While the topic is highly relevant to *Geoscientific Model Development* and the example of the Atlantic Niño—ENSO teleconnection demonstrates the tool's potential, several core issues must be addressed to meet GMD standards:

1. **Scientific Context and Justification:** While the manuscript provides a thorough description of MCA, it does not sufficiently position MCA within the broader landscape of statistical climate forecasting methods, such as Canonical Correlation Analysis (CCA). Given that CCA maximizes correlation—often the primary metric in forecast validation—it is important to clarify why MCA was chosen over CCA and in which scenarios it is preferable or less suitable. Additionally, *Spy4Cast* should be contextualized within the existing Python ecosystem for climate forecasting and analysis (e.g., XCast, xeofs, climpred), which similarly leverage xarray/dask. A discussion of the conceptual and practical advantages of *Spy4Cast* compared to these tools—such as what *Spy4Cast* offers that XCast does not—would strengthen the manuscript (see Specific Comments 3 and 8).

2. **Scalability and Computational Efficiency:** The manuscript states that *Spy4Cast* can handle ”large” climate datasets, but its reliance on in-memory NumPy arrays for computing the cross-covariance matrix raises concerns about scalability. High-resolution or global-scale datasets could result in excessive memory usage, potentially limiting the tool's applicability to gigabyte- or terabyte-scale datasets. The authors should clarify the realistic data-size limits of *Spy4Cast* and discuss potential strategies for improving scalability, such as incorporating `dask.array`, processing data in smaller subregions, or applying prior dimensionality reduction techniques like PCA (see Specific Comment 1).

3. **Scope and Flexibility:** The manuscript suggests that *Spy4Cast* can handle “any kind of predictor,” yet the current implementation appears tailored to oceanic fields with latitude-longitude-time coordinates. While the library's specialized design provides advantages, it also seems to come at the cost of flexibility—for example, through the use of custom classes such as `Dataset` and `Region`. It would be helpful to clarify whether the tool can process land-based variables, vertical levels (e.g., depth or pressure), or multiple variables simultaneously (see Specific Comments 21, 22, and 25).

4. **Clarity, Structure and Documentation:** Although *Spy4Cast* is openly licensed and includes example scripts, the documentation remains sparse, requiring users to infer key details from the source code. The manuscript refers to preprocessing steps (e.g., detrending, filtering) but does not fully explain their implementation. Additionally, some design choices—such as the use of specialized `Dataset` and `Region` classes instead of a pure xarray-based approach—should be more clearly justified, with discussion of their advantages and limitations. The structure of the manuscript would also benefit from a clearer distinction between the two validation approaches used: historical split validation and leave-one-year-out cross-validation. Finally, a more comprehensive user manual—including details on parameter options, algorithmic considerations (e.g., the use of randomized SVD), and potential pitfalls—would enhance reproducibility and usability (see Specific Comments 2,18,19,20,29,24,28, and 27).

In summary, *Spy4Cast* shows promise for moderate-scale, MCA-based forecasting research. However, its impact would be strengthened by (i) a broader discussion of MCA in relation to other statistical methods, (ii) a clear assessment of its current scope and scalability limitations, and (iii) improved presentation and documentation, either within this manuscript or through comprehensive online resources.

Specific Comments

Comment 1 (L7; also L48-53, 67, 69)

The manuscript states that *Spy4Cast* "enables large dataset manipulation," yet the internal computations appear to rely on in-memory NumPy arrays for constructing the cross-covariance matrix. To clarify the tool's practical limits, please specify:

- The maximum feasible input size (e.g. grid dimensions) before memory constraints become prohibitive.
- Whether *Spy4Cast* integrates with `dask.array` or chunked xarray computations, specifically the computation of the SVD. If not, consider explicitly stating that *Spy4Cast* is suited for moderate-scale datasets but may not efficiently handle very large (gigabyte-scale and beyond) datasets.

Additionally, please revisit the use of the term "scalability" (L69). While static type-checking improves code maintainability, it does not directly enhance the tool's ability to handle large datasets. Consider rephrasing for accuracy.

Comment 2 (L13) *. . . well documented for beginners and experience programmers.*

The statement about the documentation being suitable for both beginners and experienced programmers may need reconsideration. The current documentation appears quite minimal—while one notebook demonstrates the general workflow, there is little to no explanation of the individual preprocessing steps and their implications. Expanding the documentation to provide clearer descriptions of these steps would greatly enhance usability.

Comment 3 (L16-40)

The manuscript provides a detailed discussion of MCA, while Canonical Correlation Analysis (CCA) is mentioned only briefly. Since many operational forecasting systems rely on CCA—given that it maximizes correlation, which often aligns more directly with forecast evaluation criteria—please expand on the rationale for focusing on MCA.

Specifically, clarify the trade-offs between MCA and CCA in statistical forecasting. Additionally, consider citing or discussing more recent literature (e.g., [1]) that compares these methods.

If feasible, a brief note on whether Spy4Cast could be extended or adapted for CCA—or if it is specifically designed for MCA—would be valuable.

Comment 4 (L21 onward)

The manuscript consistently refers to one field as the "predictor" and the other as the "predictand," though MCA itself is inherently symmetric and does not imply a directional causal relationship. To avoid potential misunderstanding, consider clarifying that this terminology is used specifically in the context of forecasting rather than as a fundamental property of MCA.

Comment 5 (L21) *In this context, MCA analysis provides spatial patterns of the predictor and predictand field which are related by teleconnections.*

The text suggests that all MCA modes inherently represent teleconnections or causal relationships. Please clarify that these modes primarily capture statistical covariance and do not necessarily indicate causal links.

Comment 6 (L33) *. . . it has the advantage of being easily interpretable . . .*

Higher-order MCA modes can become increasingly difficult to interpret due to the orthogonality constraints of SVD, potentially leading to so-called "Buell patterns" (similar to PCA) [e.g., 2]. Please acknowledge this limitation when discussing the interpretation of the second or third MCA modes.

Comment 7 (L41) *. . . a new paradigm of research in climate variability studies has emerged . . .*

This statement is vague for readers unfamiliar with Cai et al. (2019). Please specify what this paradigm entails (e.g., cross-basin interactions, trans-basin teleconnections) and clarify how *Spy4Cast* relates to it.

Comment 8 (L48-53)

The manuscript would benefit from a more thorough discussion of other Python-based forecasting and dimensionality-reduction packages, such as XCast [3], climpred [4], and xeofs [5]. Please compare *Spy4Cast*'s approach, strengths, and limitations relative to these tools—for example, in-memory vs. distributed computing or a specialized feature set vs. a more general framework. Expanding this discussion would help clarify *Spy4Cast*'s positioning within the broader climate data analysis software ecosystem.

Comment 9 (L83)

The definition of the cross-covariance matrix assumes zero-mean time series. Additionally, the formulation appears to be missing a normalization factor of $1/n_t$ or $1/(n_t - 1)$ for an unbiased covariance estimation. Please clarify or correct this as needed.

Comment 10 (L89) *The information of this matrix is redundant . . .*

It would be more precise to describe this as a high degree of redundancy—or even better, as multicollinearity.

Comment 11 (L89) *. . . produce the same maps . . .*

This statement is generally incorrect. You likely mean "similar maps" rather than identical ones. The claim would only be true if two time series were exactly equal, which is unlikely in practice.

Comment 12 (L89) *Also it is a complex matrix as it takes into account all possible relations between points.*

Are you certain this includes *all* possible relationships, or only linear ones? Consider revising for accuracy. Additionally, rather than making this statement, it may be more useful to emphasize that the matrix grows rapidly in size, with dimensions $n_y x n_z$.

Comment 13 (L94) *. . . matrix of eigenvalues . . .*

The result of SVD provides a diagonal matrix containing the singular values, not eigenvalues. Please correct this terminology.

Comment 14 (L94) *. . . which represents the squared covariance fraction . . .*

Each singular value represents the covariance explained, while the squared singular values correspond to the squared covariance. The Squared Covariance Fraction (SCF) can be computed from the singular values, but it is not provided directly.

Comment 15 (L100) *. . . fraction of variance . . .*

fraction of **squared co**variance

Comment 16 (L101) *. . . which are linked by having maximum covariance.*

More precisely, the first mode is linked by maximum covariance. Subsequent modes follow these principles: (i) Mode $i + 1$ captures the maximum covariance of the remaining data after the first i modes have been removed. (ii) This is subject to the constraint that the patterns R_{i+1} and Q^{i+1} are orthogonal.

Comment 17 (L120) *. . . where n is the number of observations.*

I assume that n refers to the number of grid points or spatial locations. The term "observations" may imply a temporal scale, which does not seem to be the case here. Please clarify.

Comment 18 (L134-135) *Spy4Cast is organized in three steps: setup, preprocess and methodology. The procedural workflow is illustrated in figure 1.*

The text appears to contradict Figure 1, which explicitly separates configuration and methodology, while treating preprocessing as part of methodology (see Specific Comment 46). Please clarify or revise for consistency.

Comment 19 (L135-136)

Figure 1 presents a logical workflow, yet this structure does not appear to be reflected in the design of the software. Could you clarify the reasoning behind this discrepancy?

Comment 20 (L139) *Configuration: loading data and slicing region.*

I would argue that loading data and slicing a region are distinct from configuration. Shouldn't configuration be limited to defining metadata, with data loading and preprocessing occurring separately based on that configuration? Merging these logically different concepts seems confusing. Could you clarify the reasoning behind this design choice?

Comment 21 (L140-165)

It seems that some functionalities, such as temporal and spatial slicing, are already available in xarray. Could you clarify why you chose to introduce custom classes like `Dataset` and `Region` instead of relying on xarray's native methods?

Additionally, does *Spy4Cast* support advanced xarray features like `open_mfdataset`, which are essential for working with climate datasets distributed across multiple files?

Finally, for users who need custom preprocessing steps (e.g., weighting, detrending), can they seamlessly revert to or integrate with standard xarray workflows?

Comment 22 (L140-165)

Currently, *Spy4Cast* appears to support only 2D or 3D ocean-centric data (longitude, latitude, time). If the intention is to allow "any kind of predictor" (L9,10), please clarify whether different variables (e.g., global soil moisture) or multi-dimensional data (e.g., depth, forecast members) can be seamlessly incorporated.

Additionally, do the constraints for the predictor apply equally to the predictand?

If the current implementation is specialized (e.g., limited to ocean-only fields or requiring specific naming conventions), please make these limitations explicit in the user documentation.

Comment 23 (L144) *This method will not load the data-set into memory, because it internally uses xarray function* `xarray.open_dataset`

`xarray.open_dataset` does not inherently prevent loading the dataset into memory. The key factor is whether the `chunks` argument is specified. Please clarify this point.

Comment 24 (L168-179) *This preproccesing includes calculation of seasonal means, computation of seasonal anomalies and filtering. [...] Next, the MCA is applied, with a time linear-detrending of the data by default.*

The methods for time filtering and linear detrending require more detail:

- Is the linear detrending applied gridpoint-wise, or across the monthly or seasonal time series?
- What does "frequency filtering" entail (e.g., Butterworth filter parameters, cutoff frequencies)?
- How is missing data handled (e.g., masked land areas)?

Consider including these details in the "Preprocessing" section for clarity.

Additionally, if *Spy4Cast* is primarily designed for ocean variables, please specify how continental points and other 3D dimensions (e.g., depth or altitude) are treated.

Comment 25 (L172) *All the data can be stored in* `.npy` *format.*

Storing preprocessed data in `.npy` format may reduce compatibility with xarray/dask workflows and is less common for data exchange within the climate science community. Could you clarify the rationale for choosing `.npy` over standard formats like netCDF or Zarr, which are widely used for large-scale climate data? If this choice is primarily for internal convenience, please state that explicitly.

Comment 26 (L187) *MCA and Crossvalidation*

It would be clearer to separate MCA and cross-validation into two distinct sections. In the MCA section, please specify that you use approximate algorithms based on randomized linear algebra to accelerate singular value decomposition and provide relevant references.

Comment 27 (L190) *MCA uses test-t significance technique ...*

How is the t-test applied—one-sided or two-sided? What is the target variable? I assume it is the correlation coefficients of the homogeneous/heterogeneous correlation patterns. Additionally, please account for multiple testing when interpreting p-values or discuss how this issue is addressed.

Comment 28 (196) *...can be used to calculate other regression maps using different variables and datasets for the same period analysed (see* `mca` *and* `index_regression` *in Duran-Fonseca and Rodriguez-Fonseca (2024b)).*

Please clarify this statement in more detail. Referring readers to the `mca` and `index_regression` submodules of *Spy4Cast* without further explanation makes it difficult to understand the intended meaning without digging through the code. Providing a brief description of how these regression maps are computed would improve clarity.

Comment 29 (L211) *Arguments* `map_y` *and* `map_z` *were used to create a global regression along a larger region.*

Please elaborate on the meaning of `map_y` and `map_z`. What exactly do these arguments represent? Additionally, the phrase "global regression along a larger region" is unclear—a larger region is not necessarily global. Could you clarify what is being regressed against what?

Comment 30 (L205-225)

The manuscript would benefit from a more structured explanation of validation and cross-validation:

- I suggest consolidating both discussions into a single section, as they are closely related. While cross-validation employs classical leave-one-out cross-validation (LOO-CV), the validation approach instead partitions the dataset into concurrent (historical) time slices to account for long-term temporal autocorrelation.
- Additionally, please clarify the apparent contradiction between "Spy4Cast is not designed to assess stationarity" (L56) and "Spy4Cast is able to [...] look for non-stationary relations" (L215). Clearly defining the tool's actual capabilities and limitations in this context would improve consistency.

Comment 31 (L226,227)

I find this sentence difficult to understand (apologies, as I am not a native English speaker). Could you rephrase it to clarify the intended meaning?

Comment 32 (L229-240)

To provide broader context and strengthen the demonstration of *Spy4Cast*'s applicability, consider incorporating recent references on Atlantic Niño weakening under climate change (e.g., Crespo *et al.* [6]). This would further highlight the tool's relevance to current climate research questions.

Comment 33 (L241) *...all these features are well represented by spy4cast.*

Could you clarify what is meant by this statement? For example, earlier, you mention that the Atlantic phenomenon peaks in boreal summer, but this appears to be an a priori choice in the modeling process rather than an outcome inherently produced by the software.

Do you mean that these features can be accommodated by the researcher when using *Spy4Cast*, implying that the key strength is its flexibility? If so, please rephrase for clarity.

Comment 34 (L244) *...different variables are created.*

Please specify which variables are created. Are these evaluation metrics, MCA results, or something else? Figure 4 only displays file names, making it difficult to determine this with certainty.

Comment 35 (L244) *Spy4cast identifies the Atlantic Niño in JJA as the main mode of covariability with DJF Pacific SST anomalies.*

While *Spy4Cast* does identify the Atlantic Niño in JJA as the main mode of covariability with DJF Pacific SST anomalies, this result is largely determined by the specific choice of seasons and regions (tropical Pacific and tropical Atlantic) in the analysis. Given these constraints, what alternative outcomes could have emerged? Would the result change if the predictor region were expanded?

In the introduction, you mention that MCA can help guide more advanced prediction algorithms by identifying potential predictor regions. However, your case study presupposes prior knowledge of the predictor region. Would it be more aligned with your introduction to assume little to no a priori knowledge and let MCA reveal potential predictors (e.g., Atlantic El Niño)?

This is not necessarily a suggestion for the revised manuscript, but rather a potential idea for future exploration.

Comment 36 (L246-247) *This expansion coefficient is highly correlated with all grid points in the equatorial Atlantic, shaping the Atlantic Niño phenomenon.*

Please provide a quantitative measure to support this statement. For example, you could compute the Pearson correlation coefficient between the expansion coefficients and the Oceanic Niño Index (ONI) to quantify their similarity to ENSO.

Comment 37 (L247) *This leading mode explains almost the 60 % of the covariability*

60 % of the **squared** covariance.

Comment 38 (L249)

squared covariance

Comment 39 (L257-258) *This fact does not hold for the whole period, suggesting that the linear nature of this methodology cannot always produce accurate predictions, as there are other non-linear relationshps that the MCA is not able to capture.*

You could note that a discrepancy between different evaluation metrics (e.g., high ACC but high RMSE) often indicates a bias in the error distribution. This bias may stem from the inherent linear assumptions of MCA, which, as you correctly point out, can limit its ability to predict nonlinear extreme events.

Comment 40 (L263) *This API [. . .] has proven effective in increasing productivity and the quality of research.*

Unless there is concrete evidence supporting this claim, I would suggest softening the statement. Instead, you could say that the API has the potential to improve productivity and reproducibility in research.

Comment 41 (L265) *Spy4Cast represent the beginning of a new approach to statistical seasonal forecasting . . .*

What exactly do you mean by "new approach"? MCA itself is not new—are you referring to the use of predefined routines for analysis? If so, this is also not entirely novel, as operational forecast centers routinely employ such methods for statistical seasonal forecasting. To ensure accuracy, consider avoiding broad claims about "new approaches" unless specific evidence or metrics are provided.

Comment 42 (L267) *. . . this API is more versatile . . .*

Please clarify what makes this API "more versatile" compared to existing open-source solutions. Providing specific examples or comparisons would help substantiate this claim.

Comment 43 (L270) *Indeed, within the OFF project, it is being integrated into ESMValTool.*

Could you provide more details on the planned integration with ESMValTool, such as the expected timeline and intended functionality? Additionally, if "OFF" refers to a specific project, please define the acronym for clarity.

Comment 44 (Conclusion and Discussion)

Consider separating the discussion of the tool's limitations and future directions from the concluding remarks to follow a more standard "Conclusion and Outlook" structure.

Comment 45 (Listings 1-10)

The numerous short code listings may quickly become outdated if the API changes. Consider moving them to an online supplement or user guide while keeping only a concise set of essential examples in the main text. This would help maintain the paper's focus while ensuring comprehensive examples remain accessible in the documentation. Additionally, reassess whether all listings are necessary in their current form—Listing 2, for instance, provides limited information.

Comment 46 (Figure 1)

Could you clarify the meaning of the left-hand-side arrow? Additionally, there appears to be an inconsistency between the manuscript (L134) and Figure 1 regarding the workflow structure. The text describes three steps: configuration, preprocessing, and methodology (which includes MCA and validation). However, in the figure, the workflow is grouped into only two categories: configuration and methodology, with preprocessing included under methodology. I would argue that temporal and spatial slicing operations are also part of preprocessing.

Please explain and justify why the workflow is structured this way. Additionally, the different abstraction levels implied by the colors, boxes, and shapes are somewhat unclear. A more explicit explanation of how these visual elements correspond to the workflow's logical structure would improve clarity.

Comment 47 (Figure 4)

This figure is difficult to interpret without additional context. If it is meant as a quick reference, consider adding a brief explanation in the caption about the typical use and content of each array. Otherwise, reassess whether the figure is essential for the manuscript's long-term clarity and sustainability (cf. Specific Comment 45).

Comment 48 (Figure 5)

It appears that this figure presents the output of the MCA (expansion coefficients and homogeneous/heterogeneous correlation patterns for modes 1 to 3), yet the caption states "predicting Niño." However, no actual prediction is shown—only the covarying patterns of variability between time-lagged SST in the tropical Pacific and Atlantic.

Comment 49 (Figure 6)

When presenting ACC and RMSE, please clarify the temporal dimension over which they are computed and specify the reference variable or index. Additionally, provide units for RMSE (presumably °C).

It would also be helpful to briefly discuss the error in the context of SST variability—is the RMSE relatively low or large compared to typical SST variations?

Comment 50 (Figure 6)

Please provide a more detailed description in the caption. Specifically, what do the orange dots in the upper-right panel represent? Do they indicate uncertainties in ACC values? If so, how are these uncertainties calculated?

Comment 51 (Figure 7)

RMSE is typically non-negative. Could you clarify how a negative RMSE appears in the figure?

Comment 52 (Figures–General)

- Some figures (particularly Figures 2, 3, 6, and 7) appear to use jet colormaps. Consider using perceptually uniform alternatives, as jet can introduce visual distortions that misrepresent the data [7]. A better approach would be to match the colormap to the data type—e.g., using sequential colormaps for continuous data and diverging colormaps for anomalies. The *cmocean* package [8] provides useful options.
- Label each sub-panel clearly (e.g., A, B, C, etc.).
- Specify the plotted variables, units, and relevant domain, either within the figure or in the caption. If using shorthand notations (e.g., R, U), provide a brief explanation in the caption.
- Consider whether all figures are essential to the discussion. For example, Figure 2 (climatology) and Figure 3 (anomaly pattern) depict standard visualizations that can be easily produced with xarray. While demonstrating quick visualization is useful, these figures may not add significant value to the manuscript. At a minimum, climatology and anomaly plots could be combined into a single figure to improve conciseness.

Technical Corrections

L47: Please check the reference.

Section 4: The title seems incomplete. Consider rephrasing the section title to something like "Application: Atlantic–Pacific Teleconnections for ENSO Prediction" to more accurately describe the scope.

L140: Please check the reference. Do you mean Rew & Davis [9]?

L225: Do you mean hot topic?

L247: phenomenon

References

1. Swenson, E. Continuum Power CCA: A Unified Approach for Isolating Coupled Modes. *Journal of Climate* **28,** 1016–1030. doi:10.1175/JCLI-D-14-00451.1 (2015).
2. Richman, M. B. Rotation of Principal Components. *Journal of Climatology* **6,** 293–335. doi:10.1002/joc.3370060305 (1986).
3. Hall, K. J. C. & Acharya, N. XCast: A python climate forecasting toolkit. *Frontiers in Climate* **4.** doi:10.3389/fclim.2022.953262 (2022).
4. Brady, R. X. & Spring, A. Climpred: Verification of Weather and Climate Forecasts. *Journal of Open Source Software* **6,** 2781. doi:10.21105/joss.02781 (2021).
5. Rieger, N. & Levang, S. J. xeofs: Comprehensive EOF Analysis in Python with Xarray. *Journal of Open Source Software* **9,** 6060. doi:10.21105/joss.06060 (2024).
6. Crespo, L. R., Prigent, A., Keenlyside, N., Koseki, S., Svendsen, L., Richter, I. & Sánchez-Gómez, E. Weakening of the Atlantic Niño Variability under Global Warming. *Nature Climate Change* **12,** 822–827. doi:10.1038/s41558-022-01453-y (2022).
7. Crameri, F., Shephard, G. E. & Heron, P. J. The Misuse of Colour in Science Communication. *Nature Communications* **11,** 5444. doi:10.1038/s41467-020-19160-7 (2020).
8. Thyng, K., Greene, C., Hetland, R., Zimmerle, H. & DiMarco, S. True Colors of Oceanography: Guidelines for Effective and Accurate Colormap Selection. *Oceanography* **29,** 9–13. doi:10.5670/oceanog.2016.66 (2016).
9. Rew, R. & Davis, G. NetCDF: An Interface for Scientific Data Access. *IEEE computer graphics and applications* **10,** 76–82 (1990).