Towards Automating Time Series Analysis for the Paleogeosciences

Deborah Khider
khider@usc.edu
University of Southern California-Information Sciences Institute
Marina Del Rey, CA

Pratheek Athreya
University of Southern California-Department of Computer Science
Los Angeles, United States
pathreya@usc.edu

Varun Ratnakar
University of Southern California-Information Sciences Institute
Marina Del Rey, CA
varunr@isi.edu

Yolanda Gil
University of Southern California-Information Sciences Institute
Marina Del Rey, CA
University of Southern California-Department of Computer Science
Los Angeles, United States
gil@isi.edu

Feng Zhu
University of Southern California-Department of Earth Sciences
Los Angeles, United States
fengzhu@usc.edu

Myron Kwan
University of Southern California-Department of Computer Science
Los Angeles, United States
myronkwa@usc.edu

Julien Emile-Geay
University of Southern California-Department of Earth Sciences
Los Angeles, United States
julieneg@usc.edu

ABSTRACT

There is an abundance of time series data in many domains. Analyzing this data effectively requires deep expertise acquired over many years of practice. Our goal is to develop automated systems for time series analysis that can take advantage of proven methods that yield the best results. Our work is motivated by paleogeosciences time series analysis where the datasets are very challenging and require sophisticated methods to find and quantify subtle patterns. We describe our initial implementation of AutoTS, an automated system for time series analysis that uses semantic workflows to represent sophisticated methods and their constraints. AutoTS extends the WINGS workflow system with new capabilities to customize general methods to specific datasets based on key characteristics of the data. We discuss general methods for spectral analysis and their implementation in AutoTS.

KEYWORDS

automated machine learning, time series data, time series analysis, semantic workflows

MileTS '20, August 24th, 2020, San Diego, California, USA

ACM Reference Format:
Deborah Khider, Pratheek Athreya, Varun Ratnakar, Yolanda Gil, Feng Zhu, Myron Kwan, and Julien Emile-Geay. 2020. Towards Automating Time Series Analysis for the Paleogeosciences. In *MileTS '20: 6th KDD Workshop on Mining and Learning from Time Series, August 24th, 2020, San Diego, California, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.xxxxx

1 MOTIVATION AND OVERVIEW

Time series data is ubiquitous in many fields of research including geosciences, finance, economics, health, engineering, environmental sciences, and social sciences. Quantitative analyses in these fields use statistical methods to identify trends, patterns, and correlations within sequential data to ultimately generate signals or filters based on inference or prediction. Our work focuses on paleoclimate data, a field of climate science that focuses on understanding past climate variability. In this area, applications of time series analysis mainly focuses on signal processing. Although many signal processing methodologies relevant to time series analysis are widely available in popular packages and libraries in Matlab, Python, and R, there are important aspects that require sophisticated expertise:

(1) *Identifying methods and set parameters that are appropriate for a given dataset.* Paleoclimate datasets are often unevenly-spaced in time, requiring specific methods designed to deal with missing values such as the Lomb-Scargle Fourier Transform [24, 27, 28] and the weighted wavelet Z-transform

(WWZ, [3, 22, 33]) or hypothesizing over missing values using, for instance, interpolation or singular spectrum decomposition.
(2) *Preparing data for analysis*, including detrending, standardization, and removal of outliers.
(3) *Specifying the null hypothesis*. To evaluate the significance of spectral peaks in a climate time series, a researcher needs to generate a null hypothesis. Ideally, climatologists would want to use long runs of unforced climate simulations from numerical models. However, these series are prohibitively expensive to generate, and researchers rely on autoregressive models, most often of order 1. Depending on the complexity of these models, several parameters need to be set to simulate the behavior of climate (e.g., the autocorrelation coefficient in AR(1) models), which are either obtained from the series itself or from prior knowledge.

Paleoclimatology is an incredibly interdisciplinary field, incorporating expertise in physics, chemistry, sedimentology, biology, ecology, data science, and computer science to name a few. However, the researchers best able to understand the paleoclimate data are often not well-versed in data science and, in particular, time series analysis. Our goal is to develop automated machine learning for time series analysis. In previous work, we explored the development of automated machine learning systems for classification problems [12] and for text analytics [11, 15, 16]. A number of techniques have been developed for automated machine learning in recent years [2, 4]. These approaches search the solution space by considering different types of algorithms and exploring parameter settings, while optimizing for some target metric. Unfortunately, for time series analysis this kind of approach is unlikely to work well since it is very difficult to characterize a new time series (e.g., the noise to signal ratio, outliers, and the significance of trends in the data) from first principles without relying on proven methods that result from extensive practical experience and generalization. In addition, one must rely on synthetic series with a known behavior for approach validation. We are also interested in combining automated approaches for time series analysis with human steering, an approach that we have found effective in other contexts [8]. Finally, we aim to incorporate and compare approaches and lessons learned from different disciplines studying time series data.

Our initial work focuses on time series analysis, as a precursor to prediction tasks, and their implementation in AutoTS. Our main contribution is: (1) the compilation of expert-grade knowledge and implemented functions for spectral analysis; (2) an implementation of this knowledge in semantic workflows to enable automated generation of solutions through an intelligent workflow system; and (3) novel capabilities for workflow systems needed to support this automation, specifically data profiling and the ability to skip workflow steps.

2 APPROACH

2.1 Working with Paleogeosciences Data

Paleogeosciences group all geoscience disciplines as applied to the past, including paleoclimatology, paleoceanography, paleohydrology, paleoseismology to name a few. These disciplines all have in common the need to rely on proxy data to make inference about a past event. For instance, paleoclimatologists do not have temperature records from thermometers going back millions of years. Instead, they rely on natural archives (e.g., trees, ice cores, marine sediments) which retain information about the environment in which they were deposited. Paleoclimatologists can make measurements on these archives and interpret the climate signal. The most well-known example is the ring of trees, whose width vary with favorable (e.g., warmer, wetter conditions) vs unfavorable conditions. In this example, the tree ring width act as a proxy for past changes in temperature and precipitation.

The need for a proxy does not only apply to the environmental parameter being reconstructed but also extend to the way in which the time axis is obtained. Consider a marine sediment. Sediments accumulate on the seafloor over time, preserving a rich history of past ocean variability. Scientists obtained cores that represent this deposition history and make measurements on a finite quantity of the sediments at various depth intervals that are essentially representing time in the sedimentary archive. These sediments can be dated using the same radiocarbon techniques used to estimate the age of human settlements in archaeology, allowing to link the depth of the environmental change to time.

Because of the way paleoclimate series are obtained, signal processing is not a straightforward task:

(1) *paleoclimate time series are almost always unevenly spaced in time* since, for instance, deposition of sediments within the marine environment varies through time so that a physical 1 cubic centimeter sample may represent 100 years or 1000 years within the same archive. Although the Lomb-Scargle Fourier Transform ([24, 27, 28] and the WWZ method ([3, 22, 33]) deal with such datasets, they come with significant tradeoffs ([29]). On the other hand, interpolation can bias the statistical results and enhance the low-frequency components at the expense of the high-frequency component ([29]).
(2) *paleoclimate time series are extremely noisy*, stemming both from analytical error and weather noise.
(3) *paleoclimate time series have large uncertainties in both time and the environmental parameter*. The uncertainty arises from analytical error and the interpretation of the proxy.
(4) *paleoclimate times series display long-term, deterministic trends*, which may mask other signals of smaller magnitude that may be more relevant to the problem at hand. Therefore, detrending is often necessary. If not done carefully, the detrending can result in new, spurious signals in the analysis.

However, the ability to perform statistical analysis on paleoclimate time series is necessary to understand how our climate has changed in the past. In particular, time series analysis allows paleogeoscientists (1) to identify periodicities, which may be associated with known phenomena (e.g.[14]) and forcing (e.g., [1, 17, 20]; (2) to investigate the temporal continuum to understand how energy within the Earth system is redistributed across various timescales (e.g., [18, 25]); (3) to assign time to the long geologic record of past climate variability (e.g., [13, 23]; (4) to filter the series to highlight specific features in climate datasets (e.g., [20]); (5) to detect regime shifts (e.g., [19, 26]; and (6) to identify coherent spatiotemporal variability between multiple independent timeseries (e.g., [31]).

2.2 Spectral Analysis in the Paleogeosciences

Consider spectral analysis, a method to tease out periodic signals within time series data. In the context of paleoclimatology, it is often used to identify patterns of variability to phenomena with known periodicity such as the El Niño-Southern Oscillation or solar cycles as a basis for a mechanistic causation [1]. It has also been used to validate climate models [35]. However, paleoclimate time series data are often unevenly-spaced in time, noisy, and contains a significant amount of outliers. Furthermore, most include long-term trends reflective of known processes within the climate system (e.g., Milankovitch cycles, insolation), which are not necessarily of interest. Therefore, a common method (Fig. 1) involves significant pre-processing (standardization, detrending, removal of outliers, and hypothesizing over missing values) prior to analysis. Each pre-processing step must be performed in a particular order. For instance, interpolation results in less spurious values when performed on detrended series.

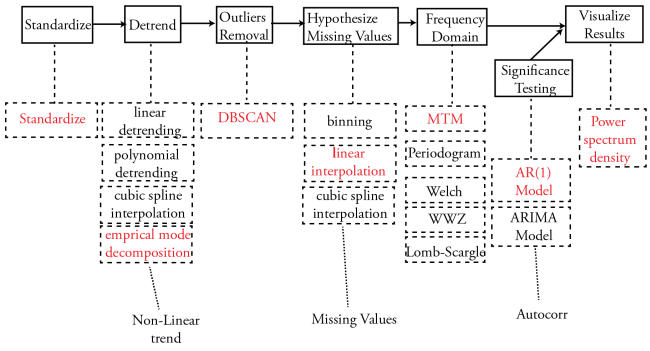

Figure 1: An expert-grade method for spectral analysis, shown at the top, includes pre-processing of the data followed by transformation into the frequency domain and significance testing. Each step can be implemented using different functions, shown in the middle. The selection of an appropriate function depends on the characteristics of the time series dataset. Functions in red illustrate such a selection. All functions are available through the Pyleoclim package [21]

In order to validate this method, we generated 720 synthetic time series with various signals, noise-to-signal ratio, trends (linear and polynomial), number of outliers and missing values, gaps, and underlying spectrum and calculated a cost function that optimizes the presence of peaks at the prescribed frequency, the width-to-height ratio of the peak (an indication of spectral leakage), and the assertion that the prescribed peaks could be significantly detected above the 95% AR(1) ensemble. The cost function was also used to search for the best default hyperparameters for each spectral method.

An example of such a validation output is shown in Fig 2. In this instance, we used the pre-processing steps highlighted in red in Figure 1 and three different analysis methods: the Lomb-Scargle transform, WWZ and multitaper method. Note that the hypothesizing missing value step was omitted for Lomb-Scargle and WWZ since these methods do not require evenly-spaced data.

Figure 2: Spectral density for a synthetic timeseries with prescribed periods of 20 and 80, a polynomial trend of order 2, a white noise/signal ratio of 0.5, 40% missing values, and outliers 5 to 7 standard deviation away from the mean using MTM, WWZ and Lomb-Scargle methods.

While the MTM and WWZ methods capture the prescribed periodicities of 20 and 80, the Lomb-Scargle method fails to detect the higher frequency signal. The Lomb-Scargle transform is known to introduce significant bias in the spectral slope [29]; but it has nonetheless been popularized within a package distributed to the paleoclimate community. This highlights the need to further validate our methods with the help of synthetic time series.

In specific cases, it is possible to validate the periodogram against the theoretical spectrum, an approach we have used with the WWZ method. This method faithfully captures the spectral slope for various underlying spectrum (red noise, colored noise, red noise with missing values, red/white noise mixture). However, since it is based on wavelets, it is fairly expensive to compute and may not be applicable to large datasets.

In the course of our validation work, we created a python package, Pyleoclim [21] that contains the core functions for spectral analysis and a series of notebooks describing the workflows. This provided the core knowledge for autoTs. In addition the Pyleoclim package contains methods for wavelet and cross-wavelet analysis, field correlation with false discovery rate, causality and (multichannel) singular spectrum analysis.

Automating this spectral analysis workflow is challenging since it requires:

- *Customized method instantiation*: Each abstract step in the template (e.g., detrending) needs to be instantiated with specific functions (e.g., linear detrending, polynomial detrending, Fig 1). The system should also allow for skipping steps depending on specific criteria, such as the spectral method not requiring evenly-spaced data and therefore the step for hypothesizing over missing values can be skipped.
- *Dynamic extraction of data characteristics*: The datasets need to be dynamically profiled to extract their characteristics as they are generated in order to decide on specializing or skipping method steps. For instance, removing outliers will result in a series with missing values after this step and, therefore, will require an interpolation step for analysis via

MTM even if the original input dataset did not contain any missing values.

2.3 Implementation in Semantic Workflows

AutoTs incorporates the methods, primitive functions, constraints, and data profilers described in Sec 2.2 and uses an intelligent workflow system to reason about all that knowledge in order to automatically generate time series analysis that are appropriate for a given dataset. AutoTs uses and extends WINGS (Workflow Instantiation and Generation System), an intelligent workflow system that uses semantic representations to capture abstract multi-step methods, and reasons about them based on constraints coming from datasets, functions, and methods [6, 10, 11]. WINGS uses Semantic Web standards (OWL, RDF, PROV) [32] for representing workflows and constraints [5]. WINGS uses AI planning algorithms, and in particular, skeletal plan refinement, to turn abstract-level workflows into fully instantiated executable workflows for data analysis. WINGS is well suited for our problem because of several existing features:

- The use of workflow templates to represent all the steps in commonly used data analysis methods [7], which we have done in many domains (e.g., [30, 34]). In this particular example, we have captured the method presented in Fig 1 as an abstract workflow template in WINGS (Fig 3).
- The use of abstract workflow components to represent abstract steps in methods.
- The use of primitive components to represent primitive functions and associate them to the abstract components [10].
- The use of metadata to capture data characteristics. When the data is uploaded in WINGS, users assert the values of the metadata properties, which can then be used to express constraints.
- The use of constraints to restrict the use of components. WINGS includes several types of constraints that represent different kinds of interdependencies across components in a workflow as well as interdependencies between components and datasets [6, 10, 11]. For instance, the invalidation rule makes the MTM component invalid in workflows where its input dataset has missing values. Other type of constraints include metadata propagation constraints and parameter setting constraints.
- The use of workflow reasoning algorithms based on AI planning that propagate constraints through a workflow and automatically generate executable workflow instances from abstract workflow templates [6].

WINGS requires two extensions in order to support the automation of time series analysis: (1) dynamic data profiling and (2) optional workflow steps, which we discuss next.

2.4 New Capabilities to Support Automation of Time Series Analysis

2.4.1 Dynamic Data Profiling. WINGS allows users to specify metadata by hand. However, AutoTS requires a mechanism to automatically run data profilers on datasets and generate metadata so that the resulting data characteristics can be used in the constraints. A data profiler is an executable that takes a dataset as input and generates a file with key value pairs that correspond to metadata

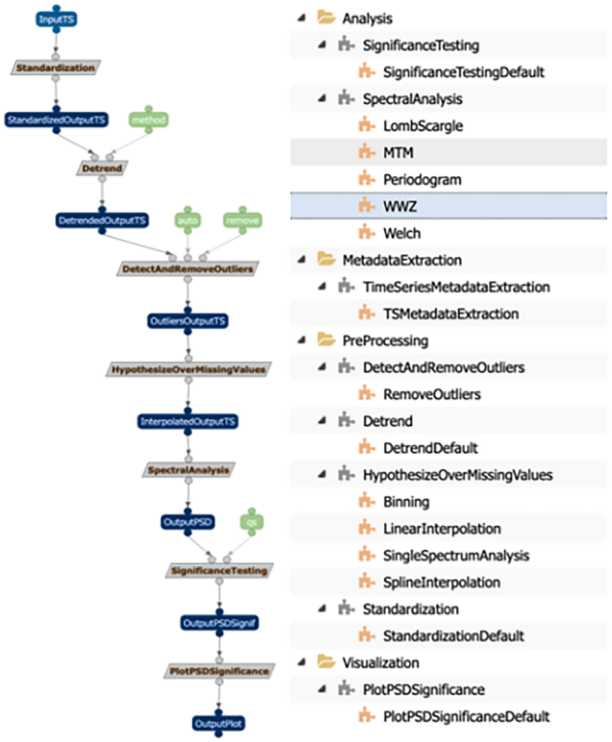

Figure 3: An abstract workflow template (left) illustrates how WINGS capture all the steps involved in the spectral analysis method shown in Fig 1. Method steps are in grey, their dataflow in blue, and parameters in green. Steps are represented in an ontology (right), where each abstract component is a class that can be instantiated with primitive components represented as subclasses.

properties extracted and their values. When a new dataset is uploaded, WINGS immediately executes the data profilers for its corresponding datatype, and uses the resulting data characteristics to assert metadata properties for the dataset.

We are extending WINGS with the ability to dynamically trigger the data profilers during execution. This will enable data profiling of new datasets generated by each workflow step, which will allow WINGS to automatically customize each step to its input dataset.

2.4.2 Optional Workflow Steps. WINGS currently requires that each and all the steps of an abstract workflow are mapped to an executable component. For AutoTS, we need to be able to decide whether or not a step in the method will be included in the final executable workflow. For instance, a HypothesizeOverMissingValues step is not needed when the spectral method is set to Lomb-Scargle or WWZ.

We created a new type of constraint in WINGS. When the constraint expression is satisfied, the abstract step is instantiated to a "no operation" step that simply passes its inputs to the next step in the workflow. For instance, a "no operation" constraint for the HypothesizeOverMissingValues step would have an expression that would trigger if the input time series is already evenly-spaced. The

requirement of being evenly spaced is a separate constraint associated with the MTM step, and that requirement is propagated to earlier workflow steps by the WINGS workflow reasoners.

Fig 4 shows two workflow instances for the general abstract workflow template shown in Fig 3. The workflow on the left uses the WWZ component, so the HypothesizeOverMissingValues step is skipped (indicated by a light color). The workflow on the right uses the MTM component, which requires its input to be evenly spaced, therefore the HypothesizeOverMissingValues step will be included.

Note that if the output of the outlier detection step is evenly spaced, then the HypothesizeOverMissingValues step should be skipped in both workflows of Figure 8. The constraints would enable this, as long as we have the metadata required for that outlier detection output once executed. Having this metadata available to check constraints during execution requires that data profiling occurs dynamically after each workflow step is executed, a capability that we are adding to the WINGS execution engine following the approach described in [9].

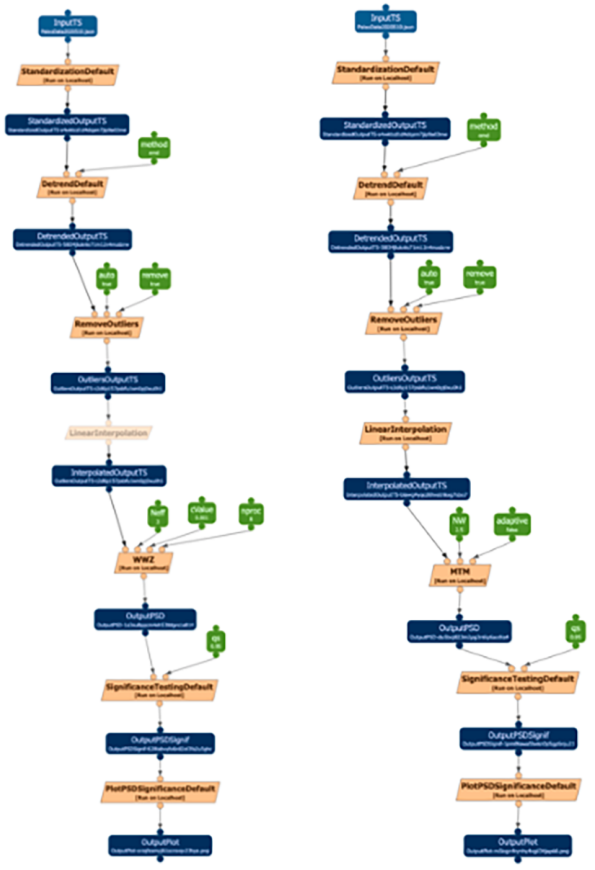

Figure 4: A step is skipped in the workflow instance on the left, because a later step (WWZ) does not need that data transformation. The workflow instance on the right includes the step because a later step (MTM) requires it.

3 CONCLUSION

We have presented a novel approach to automate time series analysis using semantic workflows that capture expert knowledge. Our main contributions are: 1) the compilation of expert-grade knowledge and implemented functions for spectral analysis, 2) an implementation of this knowledge in semantic workflows to enable automated generation of solutions through an intelligent workflow system; and 3) novel capabilities for workflow systems needed to support this automation, specifically data profiling and the ability to skip workflow steps. We have implemented this approach in AutoTS, which extends the WINGS intelligent workflow system and incorporates extensive knowledge about spectral analysis.

In future work, we plan to continue to use synthetic data to uncover new constraints about the components of spectral analysis and other time series analysis methods. While our work to date uses default hyperparemeters for each method, we plan to do parameter sweeps on our collection of synthetic series to uncover new constraints, better default parameters that take into account the entire processing pipelines and targeted defaults for different datasets. We also plan to expand our work to include other time series analysis methods, such as wavelet analysis, multi-variate series analysis, including cross-wavelet, field correlation with false discovery rate, and causality, and ultimately to do forecasting using deep learning approaches such as long short-term memory (LSTM). We are interested in testing our approach with real time series data beyond climate, particularly with financial time series data. As we gain more experience in the challenges of automating time series analysis, we will explore the role of human-guided machine learning [8] as a viable path to allow human steering of automated machine learning to incorporate intuition and creative uses of data.

4 ACKNOWLEDGMENTS

This research is funded through a grant from JP Morgan Chase Co. Any views or opinions expressed herein are solely those of the authors listed, and may differ from the views and opinions expressed by JP Morgan Chase & Co. or its affiliates. This material is not a product of the Research Department of J.P. Morgan Securities LLC. This material should not be construed as an individual recommendation of for any particular client and is not intended as a recommendation of particular securities, financial instruments or strategies for a particular client. This material does not constitute a solicitation or offer in any jurisdiction.

REFERENCES

[1] M. Debret, D. Sebag, X. Crosta, N. Massei, J. R. Petit, E. Chapron, and V. Bout-Roumazeilles. 2009. Evidence from wavelet analysis for a mid-Holocene transition in global climate forcing. *Quaternary Science Reviews* 28, 25-26 (2009), 2675–2688. https://doi.org/10.1016/j.quascirev.2009.06.005

[2] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter. 2015. *Efficient and robust automated machine learning*. Vol. 28. 2962–2970.

[3] G. Foster. 1996. Wavelets for period analysis of unevenly sampled time series. *The Astronomical Journal* 112, 4 (1996), 1709–1729.

[4] N. Fusi, R. Sheth, and H. Melih Elibol. 2017. *Probabilistic matrix factorization for automated machine learning*. ArXiv E-Prints.

[5] Daniel Garijo, Yolanda Gil, and Oscar Corcho. 2017. Abstract, link, publish, exploit: An end to end framework for workflow sharing. *Future Generation Computer Systems* 75 (2017), 271 – 283. https://doi.org/10.1016/j.future.2017.01.008

[6] Yolanda Gil, Pedro Gonzalez-Calero, Jihie Kim, Joshua Moody, and Varun Ratnakar. 2011. A semantic framework for automatic generation of computational workflows using distributed data and component catalogs. *J. Exp. Theor. Artif. Intell.* 23 (12 2011), 389–467. https://doi.org/10.1080/0952813X.2010.490962

[7] Y. Gil, P. Groth, V. Ratnakar, and C. Fritz. 2009. Expressive Reusable Workflow Templates. In *2009 Fifth IEEE International Conference on e-Science*. 344–351.
[8] Yolanda Gil, James Honaker, Shikhar Gupta, Yibo Ma, Vito D'Orazio, Daniel Garijo, Shruti Gadewar, Qifan Yang, and Neda Jahanshad. 2019. Towards Human-Guided Machine Learning. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 614–624. https://doi.org/10.1145/3301275.3302324
[9] Yolanda Gil and Varun Ratnakar. 2016. Dynamically Generated Metadata and Replanning by Interleaving Workflow Generation and Execution. 272–276. https://doi.org/10.1109/ICSC.2016.89
[10] Y. Gil, V. Ratnakar, J. Kim, P. Gonzalez-Calero, P. Groth, J. Moody, and E. Deelman. 2011. Wings: Intelligent Workflow-Based Design of Computational Experiments. *IEEE Intelligent Systems* 26, 1 (2011), 62–72.
[11] Yolanda Gil, Varun Ratnakar, Rishi Verma, Andrew Hart, Paul Ramirez, Chris Mattmann, Arni Sumarlidason, and Samuel L. Park. 2013. Time-Bound Analytic Tasks on Large Datasets through Dynamic Configuration of Workflows. In *Proceedings of the 8th Workshop on Workflows in Support of Large-Scale Science (WORKS '13)*. Association for Computing Machinery, New York, NY, USA, 88–97. https://doi.org/10.1145/2534248.2534257
[12] Y. Gil, K.-T. Yao, V. Ratnakar, D. Garijo, G. Ver Steeg, P. Szekely, R. Brekelmans, M. Kejriwal, F. Luo, and I-H. Huang. [n. d.]. P4ML: A phased performance-based pipeline planner for automated machine learning.. In *Proceedings of Machine Learning Research*, Vol. 1. 1–8.
[13] M.K. Gorman, T. M. Quinn, F.W. Taylor, J. Partin, G. Cabioch, J.A. Jr. Austin, B. Pelletier, V. Ballu, C. Maes, and S. Saustrup. 2012. A coral-based reconstruction of sea surface salinity at Sabine Bank, Vanuatu from 1842 to 2007 CE. *Paleoceanography* 27 (2012), PA3226. https://doi.org/10.1029/2012PA002302
[14] D. Gu and S.G.H. Philander. 1995. Secular changes of annual and interannual variability in the Tropics during the past century. *Journal of Climate* 8 (1995), 864–876.
[15] Matheus Hauder, Yolanda Gil, and Yan Liu. 2011. A Framework for Efficient Data Analytics through Automatic Configuration and Customization of Scientific Workflows. 379–386. https://doi.org/10.1109/eScience.2011.59
[16] Matheus Hauder, Yolanda Gil, Ricky Sethi, Yan Liu, and Hyunjoon Jo. 2011. Making Data Analysis Expertise Broadly Accessible through Workflows. (11 2011). https://doi.org/10.1145/2110497.2110507
[17] P. Huybers. 2011. Combined obliquity and precession pacing of late Pleistocene deglaciations. *Nature* 480 (2011), 229–232. https://doi.org/10.1038/nature10626
[18] P. Huybers and W. Curry. 2006. Links between annual, Milankovitch, and continuum temperature variability. *Nature* 441, 7091 (2006), 329–332.
[19] D. Khider, S. Ahn, L. E. Lisiecki, C. E. Lawrence, and M. Kienast. 2017. The Role of Uncertainty in Estimating Lead/Lag Relationships in Marine Sedimentary Archives: A Case Study From the Tropical Pacific. *Paleoceanography* 32, 11 (2017), 1275–1290. https://doi.org/10.1002/2016pa003057
[20] D. Khider, C. S. Jackson, and L. D. Stott. 2014. Assessing millennial-scale variability during the Holocene: A perspective from the western tropical Pacific. *Paleoceanography* 29, 3 (2014), 143–159. https://doi.org/10.1002/2013pa002534
[21] D Khider, F. Zhu, and J. Emile-Geay. 2019. Pyleoclim: A Python package for the analysis of paleoclimate data. https://doi.org/10.5281/zenodo.1205661
[22] J. W. Kirchner. 2005. Aliasing in 1/f(alpha) noise spectra: origins, consequences, and remedies. *Physical Review E covering statistical, nonlinear, biological, and soft matter physics* 71 (2005), 66110.
[23] L.E. Lisiecki and M.E. Raymo. 2005. A Pliocene-Pleistocene stack of 57 globally distributed benthic 18O records. *Paleoceanography* 20, PA1003 (2005). https://doi.org/1010.1029/2004PA001071
[24] N.R. Lomb. 1976. Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science* 39 (1976), 447–462.
[25] S. Lovejoy and D. Schertzer. 2013. *The weather and climate: Emergent laws and multifractal cascades*. Cambridge University Press.
[26] Eric Ruggieri and C.E. Lawrence. 2014. The Bayesian change point and variable selection algorithm: Application to the d18O proxy record of the Plio-Pleistocene. *Journal of Computational and Graphical Statistics* 23, 1 (2014), 87–110. https://doi.org/10.1080/10618600.2012.707852
[27] J.D. Scargle. 1982. Studies in astronomical time series analysis. II. Statistical aspects of spectral analyis of unvenly spaced data. *The Astrophysical Journal* 263, 2 (1982), 835–853.
[28] J.D. Scargle. 1989. Studies in astronomical time series analysis. III. Fourier transforms, aotocorrelation functions, and cross-correlation functions of unevenly-spaced data. *The Astrophysical Journal* 343, 2 (1989), 874–887.
[29] M. Schulz and M. Mudelsee. 2002. REDFIT: estimating red-noise spectra directly from unevenly spaced paleoclimatic time series. *Computers and Geosciences* 28 (2002), 421–426.
[30] Arunima Srivastava, Ravali Adusumilli, Hunter Boyce, Daniel Garijo, Varun Ratnakar, Rajiv Mayani, Thomas Yu, Raghu Machiraju, Yolanda Gil, and Parag Mallick. 2019. Semantic workflows for benchmark challenges: Enhancing comparability, reusability and reproducibility. 208–219. https://doi.org/10.1142/9789813279827_0019
[31] J.E. Tierney, J.E. Smerdon, K.J. Anchukaitis, and R. Seager. 2013. Multidecadal variability in East African hydroclimte contolled by the Indian Ocean. *Nature* 493, 389-392 (2013). https://doi.org/10.1038/nature11758
[32] W3C. 2020. Semantic Web Languages, Vocabularies, Inference, Linked Data, Query, and Vertical Applications. https://www.w3.org/standards/semanticweb/
[33] A. Witt and A.Y. Schumann. 2005. Holocene climate variability on millennial scales recorded in Greenland ice cores. *Nonlinear processes in Geophysics* 12 (2005), 345–352.
[34] C. L. Zheng, V. Ratnakar, Y. Gil, and S. K. McWeeney. 2015. Use of semantic workflows to enhance transparency and reproducibility in clinical omics. *Genome Med* 7, 1 (2015), 73. https://doi.org/10.1186/s13073-015-0202-y
[35] F. Zhu, J. Emile-Geay, N. P. McKay, G. J. Hakim, D. Khider, T. R. Ault, E. J. Steig, S. Dee, and J. W. Kirchner. 2019. Climate models can correctly simulate the continuum of global-average temperature variability. *Proc Natl Acad Sci U S A* (2019). https://doi.org/10.1073/pnas.1809959116

A ONLINE RESOURCES

The Pyleoclim package used to generate the results (including the figures) is available on GitHub (https://github.com/LinkedEarth/Pyleoclim_util/releases/tag/0.5.1beta0). A Docker container is available through DosckerHub (https://hub.docker.com/r/kcapd/pyleoclim).

The spectral workflow notebook is available on GitHub (https://github.com/KnowledgeCaptureAndDiscovery/autoTS/blob/master/notebooks/Spectral%20Workflow.ipynb)