# Functional Data Depth for the Analysis of Earth Surface Temperatures

Michele Cavazzutti, Eleonora Arnone, Ying Sun, Marc G. Genton and
Laura M. Sangalli

**Abstract** In this work, we introduce an integrated depth measure for functional data defined over complex multidimensional domains. We consider functional data whose discrete realizations are irregularly spaced, and may be available only over portions of the domain. To address this issue, we propose an integrated depth based on a Voronoi tessellation of the multidimensional domain. This approach ensures favorable statistical properties for the proposed depth, as well as computational efficiency, enabling the analysis of large-scale functional datasets. We validate our proposal with the study of air temperatures across the Earth surface, as provided by the CESM2 Large Ensemble Community Project. The proposed depth is able to capture the increase in global temperatures since the 1980s, coherently with global warming.

## 1 Introduction

The data depth is a measure of *centrality* of a datum with respect to a distribution. At first, it was introduced to establish an inward-outward ordering, and therefore

Michele Cavazzutti
Politecnico di Milano, Milan, Italy, e-mail: Michele.Cavazzutti@polimi.it

Eleonora Arnone
Università di Torino, Tourin, Italy, e-mail: Eleonora.Arnone@unito.it

Ying Sun
King Abdullah University of Science and Technology, Thuwal, KSA, e-mail: Ying.Sun@kaust.edu.sa

Marc G. Genton
King Abdullah University of Science and Technology, Thuwal, KSA, e-mail: Marc.Genton@kaust.edu.sa

Laura M. Sangalli
Politecnico di Milano, Milan, Italy, e-mail: Laura.Sangalli@polimi.it

a ranking, in the context of multivariate data [18, 9, 19]. Starting from the early 2000s [5], it was also generalized to analyze functional data [10, 11]. In addition, the depth-induced rankings have allowed the solution of various statistical problems, ranging from visualization and outlier detection [7, 16, 17] to hypothesis testing, as well as clustering and classification [9].

The majority of the existing literature on functional data depth focuses on functional data defined over one-dimensional domains. Only recently some authors have started to focus on the problem of providing a data depth for functional data defined over multidimensional domains, mainly considering simple two-dimensional planar domains and hypothesizing the functional data observed on regular uniform grids [6, 12]. However, many applications of interest concern data observed over complex multidimensional domains and with measurement locations that are irregular in space; furthermore, observations may be affected by some severe missing data patterns. These problems may be handled naturally within the framework of functional integrated depth, that has been recently extended also to partially observed functional data over one-dimensional domains [3, 4]. However, the discretization of the integrated depth needs to be handled cautiously, to guarantee favorable consistency properties for real data sample. In this work, we propose to discretize the integrated depth based on a Voronoi tessellation of the domain, which implicitly takes into account the complex geometry of the domain, and also permits to correctly treat irregular spatial measurements. The use of the Voronoi tessellation is also advantageous in terms of computational efficiency. Indeed, while the existing functional depth discretizations require the ranking (sorting) of the functional evaluations in each measurement location [1, 13, 3], we only need to rank the evaluations in the Voronoi seeds, significantly reducing the number of computations compared to the full set of locations.

We illustrate our method with the analysis of Earth's surface average yearly temperatures over the period 1850-2100, as provided by the CESM2 Large Ensemble Project [15]. The CESM2 Large Ensemble (LENS2) consists of a set of temperature simulations at a 1-degree spatial resolution, covering the period 1850-2100 under CMIP6 historical and SSP370 future radiative forcing scenarios. In this context, we employ the proposed depth to study the evolution of average annual temperatures between 1850 and 2014. Fig. 1 shows the average annual temperatures for the years 1850 and 2014. By applying the proposed depth, we observe that temperature functions have exhibited increasingly extreme behaviors since the 1980s. This finding aligns with established knowledge about global warming, and demonstrates the effectiveness of our method.

## 2 Integrated Functional Depth for Multidimensional Domains

Let $D : \mathbb{R} \times \Gamma \to [0, 1]$ be a univariate depth that satisfies the conditions expressed in [13], where $\Gamma$ is the space of distributions on $\mathbb{R}$. These conditions are required to guarantee the good properties of the overall integrated functional depth, and are

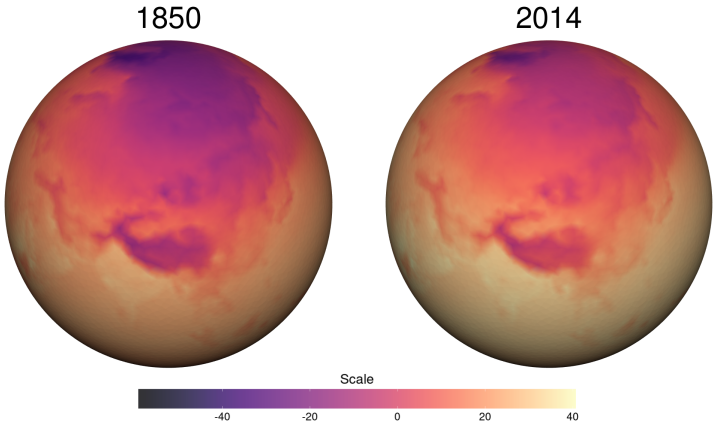


**Fig. 1** Visualization of the annual average temperature (°C) field in 1850 and 2014 from the CESM2 Large Ensemble Project

encountered by the main univariate depths of interest, such as the halfspace depth [18] and the simplicial depth [9]. Let $\Omega \subset \mathbb{R}^d$ be the support of the functional data, with $d = 2, 3$. In the application to temperature data introduced in Sec. 1, $\Omega$ is the Earth surface, a two-dimensional manifold embedded in $\mathbb{R}^3$. Consider a stochastic process $X : \Omega \to \mathbb{R}$ with law $P$. Let $P_{\boldsymbol{p}}$ be the marginal distribution of $X(\boldsymbol{p})$, $\forall \boldsymbol{p} \in \Omega$.

We consider a natural extension of the definition of the weighted integrated functional depth, originally developed for a one-dimensional domain in [1]. Let $\phi$ be a positive continuous function on $\Omega$, $\phi(\boldsymbol{p}) > 0 \; \forall \boldsymbol{p} \in \Omega$. We define the following weighting function:

$$w_\phi(\boldsymbol{p}) = \frac{\phi(\boldsymbol{p})}{\int_\Omega \phi(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}}. \tag{1}$$

We can thus define the Weighted Integrated Functional Depth (WIFD) for multi-dimensional domains as follows. The Weighted Integrated Functional Depth of the continuous function $x : \Omega \to \mathbb{R}$, with respect to $P$, is given by

$$\mathrm{WIFD}(x, P) = \int_\Omega D(x(\boldsymbol{p}), P_{\boldsymbol{p}}) w_\phi(\boldsymbol{p}) \mathrm{d}\boldsymbol{p}. \tag{2}$$

The weight function is used to weight differently different portions of the domain. For instance, in the analysis of temperature across the Earth surface, we know that the variability is reduced in water areas with respect to the inland, due to the water's thermal inertia, and we may wish to weight more either water or land areas, depending on the target of the analysis. Notice that the weight function can also be used to address partial observability in the functional data, as done by [3, 4] in the context of univariate functional data over one-dimensional domains.

## 2.1 Depth for Discrete Measurements

In practice the distribution $P$ is not known in advance, and needs to be estimated with a sample distribution. Let $\{X_i : \Omega \to \mathbb{R}\}_{i=1,\dots,n}$ be a set of i.i.d. stochastic processes drawn from the distribution of interest $P$. Let $\hat{P}$ be the sample distribution that assigns weight $\frac{1}{n}$ to each $X_i$, where $n$ is the sample size, and let also $\hat{P}_{\boldsymbol{p}}$ be its marginal distribution at $\boldsymbol{p}$. The WIFD sample estimate is obtained by plugging $\hat{P}$ in (2).

Both the population and the sample versions of WIFD require the availability of continuous functional observations. However, functional data are available only through discrete measurements, and the integral in (2) needs to be approximated with a discrete sum. In the existing literature for one-dimensional domains, this problem has so far been addressed by assuming a common, regular measurement grid [14, 12], therefore weighting each point equally. From a practical point of view, this approach is equivalent to the vectorization of the functional data, therefore discarding the information related to the spatial displacement of the measurements. When dealing with multidimensional domains, the so obtained depth would be problematic for various reasons. First, the curse of dimensionality imposes a significantly larger number of measurement locations to observe each function with minimal accuracy. Even if the functions were observed on a common set of locations — which is not always the case — the computational burden of ranking functions for every location quickly becomes prohibitive. Additionally, the measurement locations might follow a highly irregular pattern, depending on the specific problem under consideration. In such cases, assigning equal weights to all locations can disproportionately amplify the influence of areas with a higher density of observations, skewing the overall results. These factors collectively pose significant challenges in accurately and efficiently determining functional depth measures for multidimensional domains. Such difficulties can be solved by means of an appropriate discretization of the integral.

In this work, we consider a Voronoi tessellation of the multidimesional domain $\Omega$; see, e.g., [2]. Voronoi tessellations can represent accurately complex multidimensional geometries, including non-convex domains; moreover, they can be easily obtained as the dual representation of triangular meshes, as shown in Fig. 2 for a simple two-dimensional domain.

In order to approximate the depth's integral, we represent the discrete realization of the functional data with a step function, that takes as constant value, in each Voronoi cell, the average of the datum's measurements in the cell itself. Thus, the point-wise depth in (2) is substituted in each cell by the univariate depth of $n$ averages, one for each functional datum. As a result, the discrete version of (2) is the sum of the univariate depths computed in each cell multiplied by the Lebesgue measure of the cell itself.

The approach can naturally handle different measurement grids across different statical units, as well as irregular density of the measurement locations, without overemphasis on the regions of the domain with more observations. For instance, in the application to temperature data, the functions are measured on a regular longitude-latitude grid (of size 55296), leading to much denser locations near the

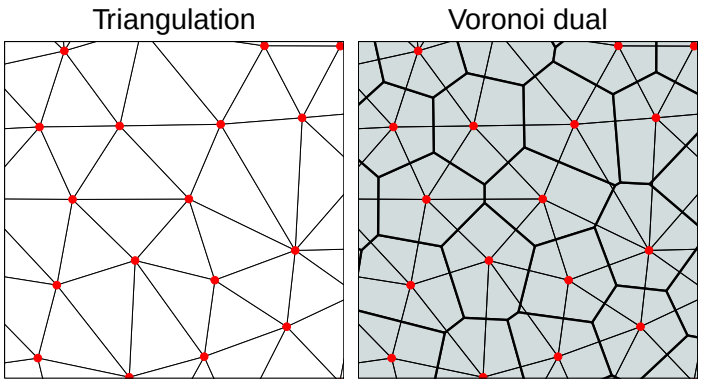


**Fig. 2** Example of domain discretization in two dimensions. In the left panel, we show a detail of the triangulation of a two-dimensional planar domain. In the right panel, we show instead the corresponding dual Voronoi tessellation, where the edges of the Voronoi cells are depicted with thicker linewidth with respect to the edges of the triangulation

poles with respect to the equator zone. Using a regular Voronoi tessellation of the globe, instead of the point-wise observational grid, we avoid over-weighting the contribution of the poles with respect to areas close to the equator. In addition, the computational cost related to this formula is now depending on the number of Voronoi cells, and no longer on the number of locations, allowing the analysis of large data sets.

## 3 Application to LENS-CESME Temperature Data

The CESM2 Large Ensemble Community Project (LENS2) [15] is a publicly accessible collection of climate model simulations designed to enhance knowledge of internal climate variability and climate change. Among the various simulations available, we are interested in analyzing the set of surface temperatures in the period 1850-2014. The data sample consists of 165 annual temperature simulations observed at 1-degree spatial resolution, obtained under CMIP6 historical radiative forcing scenarios. An example of such functions is depicted in Fig. 1. Given that global warming has become to be evident from the late 1980s, we expect the temperature functions to become more and more extreme after that time as whole functional objects across the Earth surface. Therefore, we seek for validation from our methods by applying the proposed depth to the aforementioned functional temperatures. In this context, the domain is a two-dimensional sphere embedded in a three-dimensional space. As anticipated above, the temperature surfaces are measured on a regular longitude-latitude grid with 55296 locations.

We discretize the domain using a uniform triangulation of the domain of 13969 nodes, and the corresponding dual Voronoi tessellation. We then compute the pro-

posed WIFD for the 165 functions using the univariate simplicial depth for $D$ [9]. In

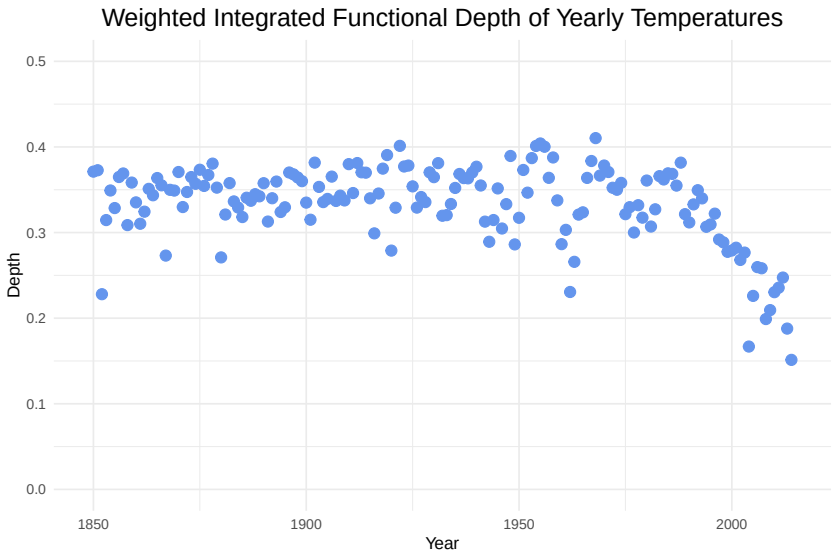


**Fig. 3** WIFD computed with point-wise simplicial depth $D$ for the Yearly Temperatures in the period 1850-2014. Notably, the depth seems to decrease with time starting from 1980s on, and the great majority of the most extreme years is clearly clustered in the last decade

Fig. 3 we show the depth associated with each annual temperature function against the year. The data shows a noticeable decrease in depth starting from the 1980s, signaling a trend of increasingly extreme air temperatures over time. This result corroborates knowledge about global warming.

## 4 Conclusion

In this work we illustrated a functional depth for data defined on multidimensional domains. Our proposal handles the difficulties related to the complexity of the domain by means of a Voronoi-based integration. We apply the proposed depth to study the global surface temperatures estimates for the years 1850-2014, showing that the climate conditions have started becoming extreme in the last three decades.

There are several directions of research that could be explored, starting from this work. One possible extension would be to exploit the weight function in (1) to handle functional data that are only partially observable, similarly to what is done in [3, 4] in the one-dimensional domain setting. Another natural extension of the proposed depth goes toward functional vectors.

## Acknowledgments

L.M. Sangalli acknowledges the PRIN 2022 project CoEnv - Complex Environmental Data and Modeling (CUP 2022E3RY23), funded by the European Union – NextGenerationEU programme, and by the Italian Ministry for University and Research. L.M. Sangalli also acknowledges the project GRINS - Growing Resilient, INclusive and Sustainable (GRINS PE00000018 – CUP D43C22003110001), funded by the European Union - NextGenerationEU programme. The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them. M. Cavazzutti and L.M. Sangalli acknowledge MUR, Dipartimento d'Eccellenza 2022-2027, Dipartimento di Matematica, Politecnico di Milano. M. Cavazzutti also acknowledges financial support from KAUST as a visiting student.

## References

1. Claeskens G., Hubert M., and Slaets L. and Vakili K.: Multivariate Functional Halfspace Depth. Journal of the American Statistical Association, **109**(505), 411-423 (2014)
2. De Berg M.: Computational geometry: algorithms and applications. Springer Science & Business Media (2000)
3. Elías A., Jiménez R., Paganoni A. M., Sangalli L. M.: Integrated Depths for Partially Observed Functional Data. Journal of Computational and Graphical Statistics, **32**(2), 341–352 (2022)
4. Elías A., Nagy S.: Statistical properties of partially observed integrated functional depths. Test, to appear (2024)
5. Fraiman R., Muniz G.: Trimmed means for functional data. Test, **10**(2), 419–440 (2001)
6. Genton M. G., Johnson C., Potter K., Stenchikov G., Sun, Y.: Surface boxplots. Stat, **3**(1), 1-11 (2014)
7. Hyndman R. J., Shang H. L.: Rainbow plots, bagplots, and boxplots for functional data. Journal of Computational and Graphical Statistics, **19**(1), 29–45 (2010)
8. Kokoszka P., Reimherr M.: Introduction to Functional Data Analysis. Chapman and Hall/CRC, New York (2017)
9. Liu R. Y.: On a notion of data depth based on random simplices. The Annals of Statistics, **18**(1), 405–414 (1990)
10. López-Pintado S., Romo J.: On the Concept of Depth for Functional Data. Journal of the American Statistical Association, **104**(486), 718–734 (2009)
11. López-Pintado S., Romo J.: A half-region depth for functional data. Computational Statistics & Data Analysis **55**(4), 1679–1695 (2011)
12. López-Pintado S., Wrobel J.: Robust non-parametric tests for imaging data based on data depth. Stat, **6**, 405–419 (2017)
13. Nagy S., Gijbels I., Omelka M. and Hlubinka D.: Integrated depth for functional data: statistical properties and consistency. ESAIM: Probability and Statistics, **20**, 95-130 (2016)
14. Nagy S. and Gijbels I. and Hlubinka D.: Weak convergence of discretely observed functional data with applications. Journal of Multivariate Analysis, **146**, 46-62 (2016)
15. Rodgers K. B., Lee S.-S., Rosenbloom N., Timmermann A., Danabasoglu G., Deser C., Edwards J., Kim J.-E., Simpson I. R., Stein K., Stuecker M. F., Yamaguchi R., Bódai T., Chung E.-S., Huang L., Kim W. M., Lamarque J.-F., Lombardozzi D. L., Wieder W. R., Yeager S. G.: Ubiquity of human-induced changes in climate variability. Earth Sysem Dynamics, 12, 1393–1411 (2021)

16. Sun Y., Genton M. G.: Functional boxplots. Journal of Computational and Graphical Statistics, **20**(2), 316–334 (2011)
17. Sun Y., Genton M. G.: Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. Environmetrics, **23**(1), 54–64 (2012)
18. Tukey, J. W.: Proceedings of the 1974 International Congress of Mathematicians.Vol. 2, Canadian Mathematical Congress, Vancouver, pp. 523–531 (1975)
19. Zuo, Y., Serfling, R.: General notions of statistical depth function. The Annals of Statistics, **28**(2), 461–482 (2000)