Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

P-spline smoothing for spatial data collected worldwide

This is the final peer-reviewed author’s accepted manuscript (postprint) of the following publication:

*Published Version:*
Greco F, V.M. (2018). P-spline smoothing for spatial data collected worldwide. SPATIAL STATISTICS, 27, 1-17 [10.1016/j.spasta.2018.08.008].

*Availability:*
This version is available at: https://hdl.handle.net/11585/647337 since: 2018-10-21

*Published:*
DOI: http://doi.org/10.1016/j.spasta.2018.08.008

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

08 July 2024

This is the final peer-reviewed accepted manuscript of:

**Greco, F., Ventrucci, M., Castelli, E., 2018. P-spline smoothing for spatial data collected worldwide. Spat. Stat. 27, 1–17.**
**https://doi.org/10.1016/j.spasta.2018.08.008**

The final published version is available online at:
https://doi.org/10.1016/j.spasta.2018.08.008

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

*This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)*

***When citing, please refer to the published version.***

# P-spline smoothing for spatial data collected worldwide

Fedele Greco[a], Massimo Ventrucci[a,*], Elisa Castelli[b]

[a]*Department of Statistical Sciences, University of Bologna, via delle Belle Arti n. 41, 40126 Bologna, Italy*
[b]*Institute of Atmospheric Sciences and Climate (CNR-ISAC), Via Piero Gobetti 101, 40129 Bologna, Italy*

**Abstract**

Spatial data collected worldwide from a huge number of locations is frequently used in environmental and climate studies. Spatial modelling for this type of data presents both methodological and computational challenges. In this work we illustrate a computationally efficient non-parametric framework in order to model and estimate the spatial field while accounting for geodesic distances between locations. The spatial field is modelled via penalized splines (P-splines) using intrinsic Gaussian Markov Random Field (GMRF) priors for the spline coefficients. The key idea is to use the sphere as a surrogate for the Globe, then build the basis of B-spline functions on a geodesic grid system. The basis matrix is sparse as is the precision matrix of the GMRF prior, thus computational efficiency is gained by construction. We illustrate the approach with a real climate study, where the goal is to identify the Intertropical Convergence Zone using high-resolution remote sensing data.

*Keywords:* smoothing, intrinsic Gaussian Markov Random field, P-spline, geodesic, ITCZ

## 1. Introduction

High-resolution spatial data collected worldwide, usually by means of remote sensing techniques, is wide-spread in environmental and climate studies: most of the statistical methods developed in modelling this kind of data use the sphere as a surrogate for the Globe. Modelling data collected at a global scale presents both methodological and computational challenges. The traditional toolkit for a spatial data modeller when dealing with geostatistical datasets and aiming to make predictions at unmonitored locations would suggest to apply kriging techniques (see, e.g., Banerjee et al. (2014)). These rely on the assumption of a smooth Gaussian Random Field (GRF), continuous

*Corresponding author: massimo.ventrucci@unibo.it

in space but only observed at a discrete set of points, any finite realization of it being generated by a multivariate Gaussian distribution. The covariance structure of this distribution is specified via a spatial covariance function. The practice is largely dominated by spatial covariances defined on Euclidean distances, such as the Matérn family, thus a preliminary step in the analysis is the projection of the 3d Cartesian coordinates (from the Earth's surface) over a 2d coordinate space. The standard choice is to work with geographic coordinates (latitude-longitude), but other mapping methods can be used. Banerjee (2005) provides a review of such mapping techniques and discusses the impact of the chosen metric on spatial prediction via kriging. The traditional toolkit outlined above presents two main difficulties when modelling high-resolution data observed over a spherical domain.

The first issue is that the process of spatial prediction needs to be coherent with the geometry of the sphere. Using a planar metric over a 2d projection is inappropriate because it generates spurious anisotropy and non-stationarity of the covariance function (Banerjee (2005)). The geodesic (aka great circle) distance, i.e. the length of the shortest path between two points over the surface of a sphere, is a natural candidate for measuring distances over a spherical domain. However, using great circle distances in a Matérn family does not necessarily guarantee a positive definite covariance (Gneiting, 2013). Banerjee (2005) used a simulation to study the behaviour of different metrics regarding estimation of the Matérn covariance parameters on a region as large as Colorado, finding a substantial impact of the chosen metric on the range of the correlation function. This means that with data collected on larger regions on Earth (e.g. the whole Globe), a biased estimation of the underlying field is to be expected to some extent, when covariance functions built on Euclidean distances are used. A large number of papers have tackled this issue by essentially proposing new models for data on a spherical domain, both in a parametric and non-parametric framework.

In the parametric setting, several papers focused on building valid stochastic processes for the sphere, see, e.g., Jun and Stein (2007); Jeong and Jun (2015); Heaton et al. (2014); Porcu et al. (2016) and references therein. The stochastic partial differential equation (SPDE) approach by Lindgren et al. (2011) has gained a lot of attention recently. This approach builds a GRF with Matérn covariance as the finite element solution of a particular SPDE, an idea that can be generalized for different types of manifolds including the sphere.

In the non-parametric setting, Wahba (1981) was first to introduce smoothing splines onto the sphere, while analysing weather data collected from a large number of stations around the world. Outside the spline realm, Di Marzio et al. (2014) presented local linear regression for spherical data, including the case of smoothing of a scalar response on a spherical predictor. Wood (2017) discusses in detail the connection between spline smoothing and thin plate splines for a sphere, pointing out that low rank smoothers are also applicable to spherical data. Although low rank smoothers allows for a reduction in the number of parameters to estimate, the main role in alleviating the computational burden is played by the sparsity of the smoothing matrix, obtained by using local basis functions, i.e. non-null over a limited domain. B-splines are local functions built upon joint polynomials connected at knots, which are applied in different contexts, such as in penalized spline (P-spline) regression (Eilers and Marx, 1996). With spatial data, bivariate B-splines over triangulations (Lai and Schumaker, 2007) provide a basis for piecewise polynomial surfaces and are used in spatial models (Lai et al., 2009; Baramidze et al., 2006). Finite Elements provide an alternative basis for piecewise polynomial surfaces over triangulations (Sangalli et al., 2013). Also, more recent proposals deal with data distributed on two-dimensional general domains using finite elements (Duchamp and Stuetzle, 2003; Ettinger et al., 2016) or non-rational B-splines basis (Wilhelm et al., 2016).

The second difficulty concerning the application of kriging techniques to high-resolution global datasets is purely computational. Continuous covariance functions used in geostatistics involve a dense covariance structure for the underlying GRF. When the number of data locations $n$ is large, this modelling framework becomes impractical because of the need to invert large dense matrices, with a computational cost increasing by cubic growth with $n$. Statistics literature on the *big n problem* has boomed in the last decade, mostly due to the increasing availability of high resolution remote sensing data for environmental studies. Some of the models for large data that can be implemented in a fully Bayesian hierarchical setting (for a review see Banerjee (2017)) are based on a low-rank representation of the field (Wikle and Cressie, 1999; Banerjee et al., 2008). Other proposals look to find a sparse representation of the covariance, like in tapering (Furrer et al., 2006), or of the precision matrix (Rue and Held, 2005). In this framework, the paper by Lindgren et al. (2011) derives an approximated solution to the SPDE in terms of a Gaussian Markov Random Field (GMRF), instead of a GRF, in order to gain computational efficiency. A recent approach that allows to deal with GRFs in a computationally efficient manner is Datta et al. (2016), where sparsity is introduced without the need for dimension reduction. The fully Bayesian framework presented in this paper follows both directions, in the sense that it is built on a low-rank representation using local B-splines and exploits the sparsity induced by a GMRF prior.

We propose a computationally efficient non-parametric approach to estimate the spatial field underlying data on the sphere that properly accounts for geodesic distances between locations. Our method is based on a low-rank P-spline smoother to gain flexibility w.r.t. parametric models. The main contribution of this work is the extension of the P-spline model for smoothing data collected over a spherical domain. The model is built on a set of bivariate B-splines computed on a Geodesic Discrete Global Grid (GDGG) system (Sahr et al., 2003), yielding a quasi-regular triangular mesh over the Globe. Geodesic grids have been used in spatial statistics to create flexible multi-resolution models implemented in a likelihood-based inferential framework (Cressie and Johannesson, 2008; Nychka et al., 2015). In contrast to the latter works, in this paper we follow a fully Bayesian approach and fit the model using an efficient Gibbs sampler, exploiting sparsity of the basis matrix and of the precision of the GMRF prior. We illustrate the method on a real climate study, where the goal is to identify the *Intertropical Convergence Zone* (ITCZ) from high-resolution remote sensing data collected worldwide over sea, with missing data occurring over land.

The rest of the paper is organized as follows. Section 2 describes the dataset and application goals. Section 3 presents our proposal for smoothing data over the sphere that we dub *Geodesic P-splines*. Section 4 illustrates the method used on a climate-related case study, focusing on the detection of the ITCZ. A discussion is provided in Section 5.

## 2. Motivating example

Our interest in geodesic P-splines is motivated by a climate-related case study aimed at investigating the location of the ITCZ using satellite data. The ITCZ is a region of the atmosphere broadly located within the tropical belt where the north-east and south-east trade winds converge, an area characterised by high cloudiness and severe convective precipitation (Holton and Hakim, 2013). An important aspect regards seasonal variability in the ITCZ position: the ITCZ is located roughly North of the equator in the boreal Spring and Summer, while it migrates to southern regions in Autumn and Winter. The location of the ITCZ affects duration and intensity of the wet and dry seasons at the tropics and plays a key role in the general circulation of the atmosphere: assessing its variability is crucial for improving global climate models. Moreover, understanding the long-term trend characterizing this phenomenon is crucial for monitoring changes in climate patterns on a global scale.

The phenomenon regulating the ITCZ behaviour cannot be measured directly, hence several studies have investigated it using some suitable proxy variables, like maximum precipitation (Zhang, 2001), wind field (Žagar et al., 2011), vorticity and reflectivity of the clouds (Waliser and Gautier, 1993). As a general feature, all these studies benefit from the increasing availability of satellite measurements. In this paper we focus on data from the infrared channels of the Along Track Scanning Radiometer (ATSR) series of instruments, which were in orbit from 1991 to 2012 for the accurate retrieval of sea surface temperature. Recently, in the frame of the European Space Agency ATSR Long Term Stability project (https://earth.esa.int/web/sppa/activities/multi-sensors-timeseries/alts/about), Casadio et al. (2016) developed the Advanced Infra-Red Water Vapour Estimator algorithm (AIRWAVE) for the retrieval of the Total Column of Water Vapour (TCWV) from the ATSR measurements. In this work we use TCWV as a proxy variable for locating ITCZ.

Data on TCWV regarding year 2008 was provided by the National Research Council - Institute of Atmospheric Sciences and Climate (CNR-ISAC), Italy. Data comes as monthly averages of TCWV in a raster of dimension 360 columns (longitude values) by 180 rows (latitude values), thus each cell covers one degree over latitude and longitude. In Figure 1, the data for January and July is displayed. The AIRWAVE algorithm provides reliable data over the sea and in clear sky conditions, thus TCWV observations over land are missing (roughly a third of the total number of cells), except in areas covered by lakes. The percentage of raster cells with missing observations is about 40%.

The application goal is to estimate the ITCZ position and its uncertainty. We consider the TCWV data on the fine raster grid as point-level data observed at the centroid of each cell. We are actually managing raster data as point data. This is a standard procedure when modelling high resolution raster data, particularly when adopting splines that need to be evaluated at fixed points. The same rationale has been adopted in Eilers et al. (2006); Lee and Durbán (2009); Ugarte et al. (2012). We focus on modelling the latent field of TCWV separately for each month, deferring spatio-temporal modelling to future work. The statistical challenges we tackle in this paper are related to efficient smoothing of large data to remove measurement error and to allow for rapid predictions at unmonitored locations. We believe that the extension of Bayesian P-Splines to a spherical domain can be a valuable strategy because of its efficiency and computational stability. Bayesian inference provides immediate tools for ITCZ location, by analysing the joint posterior distribution of the latent field. One issue concerning ITCZ detection is that there is no definition in terms of a fixed threshold. This situation calls for methods to search for peaks in the latent field, bearing in mind that the ITCZ is expected to be located at the Equator.

In Section 4, the ITCZ detection problem is addressed by searching for the latitudes where the TCWV latent field shows the highest values. We provide a graphical output, by plotting the posterior probability that a point on the Earth belongs to the ITCZ.

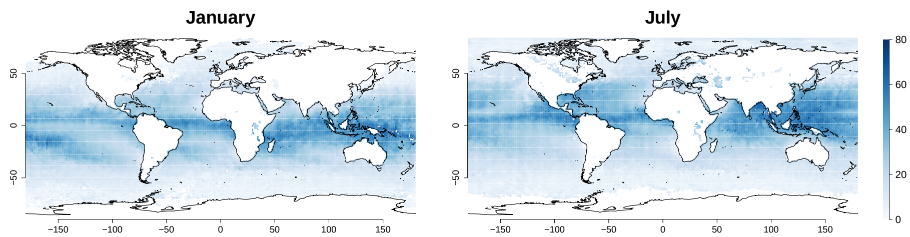


Figure 1: TCWV data for January and July (unit measure, $Kg/m^2$). In general, TCWV measurements are available only over sea, as the data cannot be accurately retrieved over land; however, note that observations are still present in correspondence of wide lakes, e.g. the Great Lakes of North America and the Victoria lake.

## 3. Geodesic P-splines

### *3.1. Background on P-splines for spatial data*

In the one-dimensional setting, P-splines (Eilers and Marx, 1996) are usually adopted to model the smooth effect of a covariate on the response as a linear combination of B-splines scaled by spline coefficients. Key features of this method are (a) equally-spaced univariate B-splines of a certain degree $d$, these being non zero over a limited interval of the covariate domain, and (b) a penalty on the $r^{th}$ order differences between adjacent spline coefficients to control smoothness. The popularity of P-splines is due to numerical stability and flexibility in the modelling choices; e.g., the penalty order and the degree of the B-splines can be decided according to the application at hand. Higher-dimensional smoothers, suitable for modelling spatial data, can be constructed as tensor product P-splines (Eilers et al., 2006). In a frequentist framework, estimation is obtained via penalized maximum likelihood or iterative re-weighted least squares, with the smoothing parameter selected via cross validation or optimized over some information criterion. This method has become increasingly popular and is currently implemented in `R` packages such as `mgcv` (Wood, 2017).

In order to build the ground for our proposal we next revise spatial P-splines for data observed on a two-dimensional latitude-longitude plane following Eilers et al. (2006). Let us assume $y_i$ is a Gaussian observation at location $(lat_i, lon_i)$, $i = 1, ..., n$, the model is

$$y_i = \mu(lat_i, lon_i) + \epsilon_i \quad ; \quad \epsilon_i \sim \mathcal{N}(0, \tau_\epsilon^{-1}),$$

where $\mu(lat_i, lon_i)$ is a two-dimensional function, with no parametric assumptions on it and $\tau_\epsilon$ is the noise precision. We can think of $\mu(lat_i, lon_i)$ as a smooth surface representing the latent field which is modelled as a linear combination of bivariate B-spline basis functions:

$$\mu(lat_i, lon_i) = \sum_{q=1}^{Q} \sum_{l=1}^{L} b_l(lat_i) b_q(lon_i) \beta_{l,q},$$

where $b_l(lat_i)b_q(lon_i)$ is the tensor product of marginal B-splines, evaluated at $(lat_i, lon_i)$, and $\beta_{l,q}$ is the associated spline coefficient. The marginal B-splines $b_l, l = 1, ..., L$ $(b_q, q = 1, ..., Q)$, are defined on a set of knots that are chosen to be equally-spaced over the latitude (longitude) domain. Taking the tensor product of the marginal bases comes to $K = QL$ bivariate B-splines built on a regular grid over the plane; see Figure 2, left panel. In this sense, P-splines give a low-rank representation of the latent field, as $K$ is typically chosen to be much lower than $n$. In matrix notation, $\boldsymbol{\mu} = \boldsymbol{B}\boldsymbol{\beta}$, where $\boldsymbol{B}$ is a basis matrix of dimension $n \times K$ and $\boldsymbol{\beta}$ the vector of spline coefficients. When data is organized in a regular grid with no missing values, the basis matrix can be computed by the Kronecker product $\boldsymbol{B} = \boldsymbol{B}_{lat} \otimes \boldsymbol{B}_{lon}$. When data are irregularly scattered over the plane, efficient row-wise Kronecker operations can still be used to compute $\boldsymbol{B}$, as this is equivalent to having data organized on a fine regular grid with missing values. We suggest the reader see Eilers et al. (2006) for details on P-splines for spatial data and to Lee (2010) for insights into the mixed model formulation of P-splines within a spatio-temporal setting.

P-splines have been framed in a fully hierarchical Bayesian context by Lang and Brezger (2004). The hierarchical model can be cast starting from the following likelihood:

$$\boldsymbol{y}|\alpha, \boldsymbol{\beta}, \tau_\epsilon \quad \sim \quad \mathcal{N}(\boldsymbol{\mu}, \tau_\epsilon^{-1}\boldsymbol{I}) \quad ; \quad \boldsymbol{\mu} = \alpha\mathbf{1} + \boldsymbol{B}\boldsymbol{\beta}$$

The penalty is reproduced by an $r^{th}$ order random walk (RW) prior on the spline coefficients, that in general can be expressed as

$$\pi(\boldsymbol{\beta}|\tau_\beta) = (2\pi)^{-\mathrm{rank}(\boldsymbol{R})/2}(|\tau_\beta \boldsymbol{R}|^*)^{1/2} \exp\left\{-\frac{\tau_\beta}{2}\boldsymbol{\beta}^\top \boldsymbol{R}\boldsymbol{\beta}\right\}, \tag{1}$$

where $\tau_\beta$ is a scalar precision hyper-parameter and $\boldsymbol{R}$ is the structure matrix of dimension $K \times K$. The non-zero entries in $\boldsymbol{R}$ impose conditional dependencies among the spline coefficients, thus encoding the type of penalty. Formally, the RW is a particular type of Intrinsic Gaussian Markov Random Field (IGMRF). The smoothing properties of an IGMRF are determined by the pattern of non-zero entries of $\boldsymbol{R}$ and by its rank deficiency. Any vector in the null space of $\boldsymbol{R}$ can be added to $\boldsymbol{\beta}$ and density (1) remains unchanged. For this reason, IGMRF priors are appropriate to model local deviations around an overall mean or, in general, a polynomial trend, with $\tau_\beta$ controlling the size of such deviations. For spatial smoothing, we will focus on a prior that leaves the overall mean unspecified, therefore $\texttt{rank}(\boldsymbol{R}) = K - 1$.

The precision matrix for P-spline smoothing over a plane proposed in Eilers et al. (2006) is constructed as the Kronecker sum

$$\boldsymbol{R} = (\boldsymbol{I}_L \otimes \boldsymbol{R}_{lon}) + (\boldsymbol{R}_{lat} \otimes \boldsymbol{I}_Q) \tag{2}$$

where $\boldsymbol{R}_{lat}$ and $\boldsymbol{R}_{lon}$ are the (marginal) structure matrices of a RW on latitudinal and longitudinal knots, respectively. If we take $\boldsymbol{R}_{lat}$ and $\boldsymbol{R}_{lon}$ as the structure of a $1^{st}$ order RW, this is equivalent to assuming an intrinsic Conditional Autoregressive (ICAR) model (Besag, 1974), with structure

$$R_{ij} = \begin{cases} k_i & i = j \\ -1 & i \sim j \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $k_i$ is the number of knots adjacent to the $i^{th}$ knot; e.g. $k_i = \{2, 3, 4\}$ according to whether $i$ is a knot on the vertex, the border, or the interior of the regular grid. Usually an ICAR prior is assumed on a set of $n$ random effects, one for each data location, but here the ICAR is on the spline coefficients. In this sense, the basis $\boldsymbol{B}$ allows the stochastic field on the $K$ spline coefficients to be expanded at a much larger number of locations like $n$. This strategy allows for a substantial reduction in the number of parameters to estimate. Choosing higher order random walks in each dimension is possible: this will yield an higher-order IGMRF prior, having a structure matrix with larger rank-deficiency; e.g. taking a $2^{nd}$ order RW on latitude and longitude comes to an IGMRF that models deviations from a plane. For a discussion of the properties of IGMRFs and their applications see Rue and Held (2005).

### *3.2. P-splines on Geodesic Discrete Global Grid Systems*

The assumption of equally-spaced knots is convenient for building Bayesian penalized spline models, because it allows to create a suitable smoothing prior by simply using an IGMRF model for regularly spaced locations on the spline coefficients. Following this idea, knot placement must take into account the geometry of the data's support. Thus, building an equally spaced basis on the latitude-longitude plane is not a sensible choice when the data covers the whole Globe or a large region thereof. Figure 2 highlights that equally spaced B-splines in terms of Euclidean distances over the latitude-longitude plane (left panel) are not equally-spaced over the sphere (right

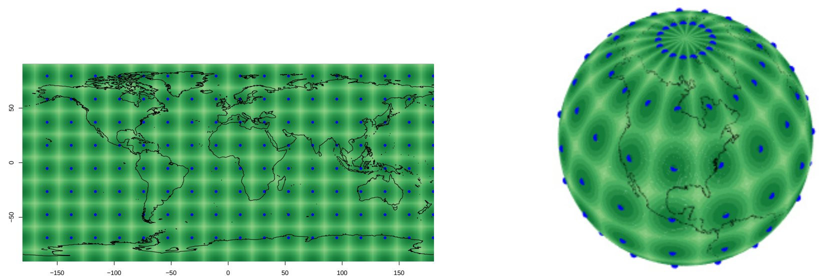


Figure 2: Cubic B-splines equally-spaced in terms of Euclidean distances over the latitude and longitude plane (left panel; computed as the tensor product of marginal B-spline basis, see Section 3.1). The right panel displays how these bases appear on the sphere.

panel). The spacing between the knots and the shape of the basis varies substantially latitude-wise: in such a knot-grid, imposing an IGMRF with structure (3) and a single precision parameter $\tau_\beta$ on the spline coefficients would generate the spurious anisotropy discussed in Banerjee (2005). Of course, this would be a naive approach to spatial smoothing over the sphere, since it does not introduce conditional dependence between knots located at extreme longitudes, which are actually close on the sphere surface. A circular penalty imposing conditional correlations among these knots seems a more sensible choice, however the irregular knot distribution over the sphere would still generate spurious non-stationarity, as this paper will discuss at a later moment. In what follows we propose an approach for (a) building geodesic knot-grids which are almost equally spaced in terms of geodesic distances, (b) building basis functions and penalty matrices on such grids.

### *3.2.1. Building the geodesic grid*

Although building *exactly* equally spaced grids over the sphere surface is an impossible task, GDGGs offer a close approximation to equal spacing and their architecture provides immediate solutions to build basis functions and penalty matrices. Details on the spatial configuration of GDGGs can be found in Randall et al. (2002). Sahr et al. (2003) outline five design choices that need to be undertaken for GDGGs construction: our choices are listed below.

1. Choice of a *base regular polyhedron*: we choose the icosahedron, which is a polyhedron made of 20 equilateral triangles and 12 nodes and consider this as a rough representation of a unit sphere. An icosahedron is displayed in Figure 3, left panel.

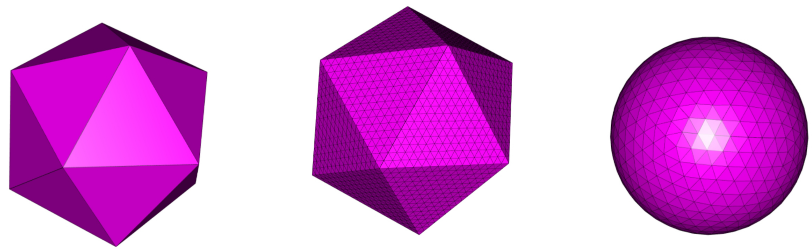

Figure 3: On the left panel, the icosahedron. On the central panel, the icomesh, i.e. the regular triangular mesh after the *split* operation is repeated four times ($\nu = 4$). On the right panel, the icosphere, i.e. the mesh obtained from normalizing the icomesh nodes of the central panel.

2. Choice of a fixed *orientation* for the base regular polyhedron relative to the Earth: we set one node of the icosahedron at coordinates $(0, 0, 1)$, assuming this to be the North Pole.
3. Choice of a *hierarchical spatial partitioning* method defined symmetrically on each face of the base regular polyhedron. At this step, we split each triangle of the icosahedron in four equal triangles. By repeating this operation an arbitrary number of times we obtain a refined mesh, which we denote as *icomesh*. In Figure 3, central panel, the reader can see the icomesh resulting from four split iterations.
4. Transforming the base polyhedron partition into the corresponding spherical surface. This is achieved by simply normalizing the icomesh nodes, so that they lay on the sphere; we denote this mesh as *icosphere*, see Figure 3, right panel. The icosphere is a refined icosahedron, hence a much better representation of the sphere.
5. Choice of a method to *assign points to grid cells*. The ability to assign points to the grid cells composing the tessellation can be useful for several purposes. In our case, it is fundamental for determining which triangle each data location falls into when it comes to the computation of the basis functions, as discussed in the following section.

Following the above five steps, we obtain a geodesic grid of knots which are almost equidistant in terms of great-circle distances. To summarize, the GDGG is constructed by splitting each icosahedron face into four triangles, in a recursive way. Note that, while the icosphere is a sphere tessellated into spherical triangles, the icomesh is a regular mesh made by equilateral triangles.

*3.2.2. Building the basis and the penalty matrix*

The number of split iterations determines the dimension of the basis, i.e. the number of columns of the basis matrix $\boldsymbol{B}$. Let $n$ be the number of observations and $\nu$ be the number of split iterations, the basis has dimension $n$ x $K$, with $K = 5 \cdot 2^{(2\nu+1)} + 2$ (Randall et al., 2002). For $\nu = 0$ we have $K = 12$, which is the number of vertices of the icosahedron, by increasing $\nu$ we obtain an icomesh with higher resolution. We adopt B-spline basis functions centred at the knots, each basis spanning six triangles (thus assuming the six closest nodes as neighbours) except for those centred at the 12 icosahedron vertices (that have five neighbours).

The next step consists of evaluating the $K$ B-splines, of a certain degree $d$, at an arbitrary data point on the sphere. Once that the triangle containing such a point is determined, B-splines can be evaluated using Bernstein polynomials (Lai and Schumaker, 2007). To this aim, we find it convenient to work on the icomesh instead of the icosphere, as it is simpler to deal with planar than with spherical triangles. Therefore, we first project the 3d data location from the icosphere onto the icomesh domain, obtaining a point, $\boldsymbol{v}$, which falls inside a planar triangle (that lies on one of the icosahedron faces) and, second, we evaluate the $K$ B-splines at this 2d point. Following Lai and Schumaker (2007), any point $\boldsymbol{v} = (x, y)$ inside a triangle of vertices $\boldsymbol{v}_1 = (x_1, y_1), \boldsymbol{v}_2 = (x_2, y_2), \boldsymbol{v}_3 = (x_3, y_3)$ has a unique representation as

$$\boldsymbol{v} = \boldsymbol{v}_1 b_1 + \boldsymbol{v}_2 b_2 + \boldsymbol{v}_3 b_3,$$

where $(b_1, b_2, b_3)$ are called barycentric coordinates and are such that $b_1 + b_2 + b_3 = 1$. The Bernstein polynomial of degree $d$ is

$$H^d_{tjk} = \frac{d!}{t!\,j!\,k!} b_1^t b_2^j b_3^k \tag{4}$$

with $t, j, k$ integer numbers summing to $d$. The following property

$$\sum_{t+j+k=d} H^d_{tjk} = 1$$

guarantees that for each location on the sphere the basis functions add up to 1. This is a desirable property for any smoothing model, giving a flat spatial field when there is no variation around the overall level, i.e. all spline coefficients are equal.

Let $z_i = (z_{i1}, z_{i2})$ denote the location for observation $i$ projected on the icomesh, $\boldsymbol{B}[i, ]$ the row entry of $\boldsymbol{B}$ with the B-splines evaluated at $z_i$ and $\{k_1, k_2, k_3\}$ the indices for the three knots closest to $z_i$ (note that these are the vertices of the triangle containing

| $d = 1$ | $d = 2$ | $d = 3$ |
|---|---|---|
| $\begin{pmatrix} H^1_{100} \\ H^1_{010} \\ H^1_{001} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$ | $\begin{pmatrix} H^2_{200} + H^2_{100} \\ H^2_{020} + H^2_{010} \\ H^2_{002} + H^2_{001} \end{pmatrix} = \begin{pmatrix} b_1^2 + b_1 \\ b_2^2 + b_2 \\ b_3^2 + b_3 \end{pmatrix}$ | $\begin{pmatrix} H^3_{300} + H^3_{200} + H^3_{100} \\ H^3_{030} + H^3_{020} + H^3_{010} \\ H^3_{003} + H^3_{002} + H^3_{001} \end{pmatrix} = \begin{pmatrix} b_1^3 + b_1^2 + b_1 \\ b_2^3 + b_2^2 + b_2 \\ b_3^3 + b_3^2 + b_3 \end{pmatrix}$ |

Table 1: Non-zero elements of $\boldsymbol{B}[i,]$, for B-splines of degree $d = \{1, 2, 3\}$.

observation $i$). It is important to note that only the B-splines centred at $\{k_1, k_2, k_3\}$ are non-zero at $z_i$, whereas the B-splines centred at the remaining knots in the icomesh are zero at $z_i$. The three non-zero elements of $\boldsymbol{B}[i, \{k_1, k_2, k_3\}]$ can be expressed as Bernstein polynomials (4), i.e. polynomials in the barycentric coordinates. Table 1 reports the non zero elements of $B[i,]$ for linear ($d = 1$), quadratic ($d = 2$) and cubic ($d = 3$) B-splines.

The resulting basis matrix $\boldsymbol{B}$ is sparse because the B-splines are non-zero over a domain spanning over only six triangles on the icomesh. Figure 4, left panel, shows how the new basis functions appear when projected over latitude and longitude. This plot suggests that a fairly similar degree of smoothness is applied everywhere using this new basis, avoiding the kind of spurious anisotropy introduced by the basis in Figure 2. The Geodesic P-splines setting is completed by specifying the matrix $\boldsymbol{R}$, that we choose as the ICAR structure (3) with rank-deficiency 1. The number of neighbouring knots is $k_i = 5$, if $i$ is one of the 12 nodes of the icosahedron, and $k_i = 6$, if $i$ is one of the remaining $K - 12$ nodes.

### *3.2.3. Model properties*

When using IGMRF priors with precision matrix $\tau_\beta \boldsymbol{R}$ on the spline coefficients $\boldsymbol{\beta}$, the structure of conditional dependence imposed by $\boldsymbol{R}$ determines the structure of the marginal variances of each coefficient, $Var(\beta_i) = \tau_\beta^{-1} R^-_{ii}, i = 1, \ldots, K$, $\boldsymbol{R}^-$ being the generalised inverse of $\boldsymbol{R}$. Different structures can lead to extremely different marginal variances. To overcome this problem, Sørbye and Rue (2014) suggest scaling the precision matrix so that the hyperprior for $\tau_\beta$ can be selected to give the same degree of smoothness, a priori, starting from different structure matrices. The scaled precision matrix can be obtained as $\boldsymbol{R}^* = \kappa \boldsymbol{R}$, where $\kappa$ is the geometric mean of the diagonal entries of $\boldsymbol{R}^-$. IGMRFs with scaled precision matrices, although being characterised by a different correlation structure, have a common feature: the average marginal variance is equal to one.

Figure 5 compares the marginal variances for three models corresponding to a naive penalty (top-left), longitude-wise circular penalty (bottom) and a geodesic penalty (top-right). For the sake of comparison, the precision matrices associated with the three

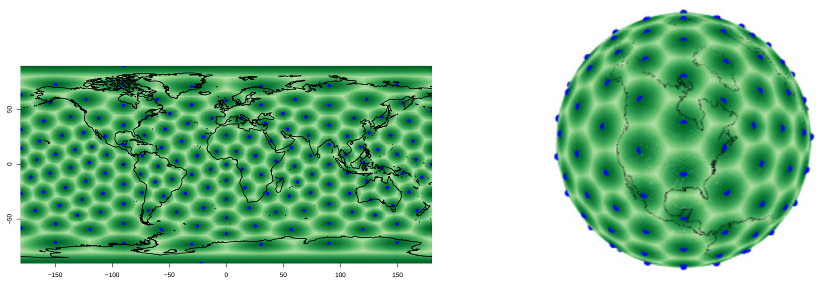


Figure 4: Cubic B-splines equally-spaced in terms of geodesic distances over the sphere (right panel; computed using Bernstein polynomials on a GDGG, see Section 3.2). The left panel displays how these bases appear on the latitude longitude plane.

models were scaled. For naive penalty, we mean an IGMRF prior for the spline coefficients laying on a planar grid, using the ICAR structure (3). The longitude-wise circular penalty is an IGMRF on a planar grid with structure (2), but assuming $\boldsymbol{R}_{lon}$ as the structure of a circular $1^{st}$ order RW. For geodesic penalty, we mean an IGMRF on a GDGG using the ICAR structure as described in Section 3.2.2. In the left panel, the non-stationarity in the marginal variances implied by using the ICAR structure on a regular planar grid (naive penalty) is evident. In the bottom panel, marginal variances obtained by building a circular penalty longitude-wise show a variation latitude-wise as expected. The IGMRF prior on the geodesic grid with the ICAR structure implies stability in the marginal variances that is not achieved with the other specifications. As a matter of fact, the geodesic grid is almost a torus since all knots, except the icosahedron nodes, have six neighbours; we believe this is a desirable feature of our model as it mimics the idea of second-order stationarity typical of Matérn correlation functions.

#### *3.2.4. Hyperpriors*

To complete the fully Bayesian model we need to set priors for the hyper-parameters $\tau_\beta$ and $\tau_\epsilon$. The precision $\tau_\beta$ regulates the amount of smoothing. When $\tau_\beta$ goes to infinity, $\boldsymbol{\mu}$ is a constant (because the rank deficiency of $\boldsymbol{R}$ is 1), while $\tau_\beta \in (0, +\infty)$ gives a more flexible surface. A standard approach is to use a Gamma, $\mathrm{Ga}(a, b)$, with shape $a$ and rate $b$, for both random walk and noise precisions. Usual parametrizations are $a$ equal to 1 and $b$ small (e.g. $\mathrm{Ga}(1, 5e{-}5)$), or $a$ and $b$ small (e.g. $\mathrm{Gamma}(1e{-}3, 1e{-}3)$), as an attempt of non informativeness on the variance scale. Several papers in the literature have discussed issues related to the Gamma conjugate priors in hierarchical additive models and proposed alternatives (Gelman, 2006; Simpson et al., 2017). Typically, the main impact regards the prior for the random walk precision, whereas the

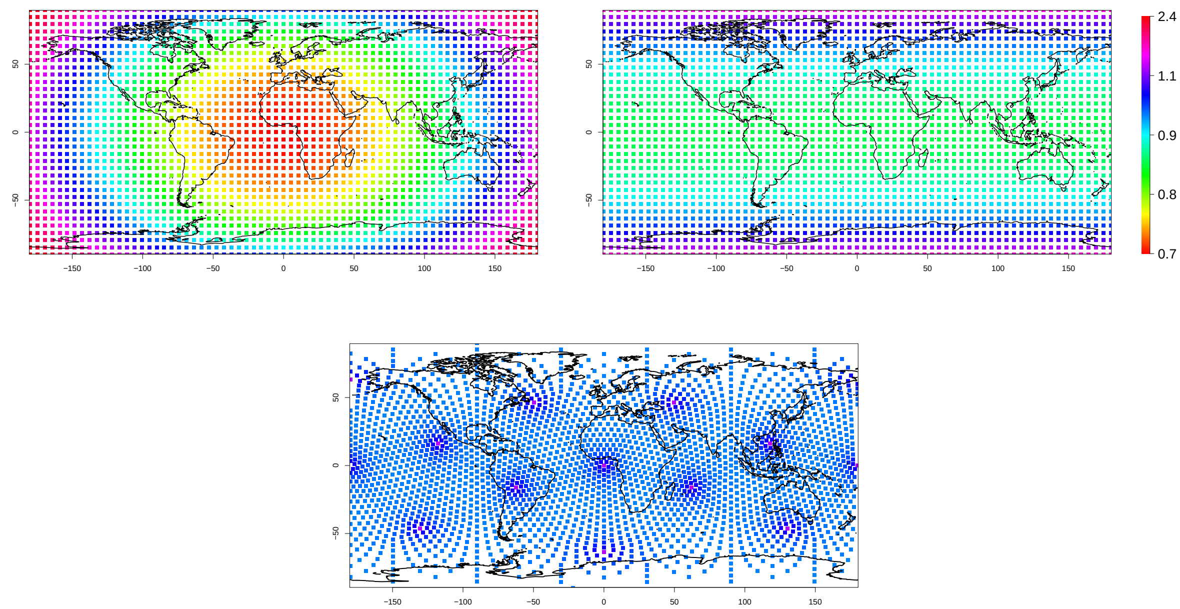


Figure 5: Marginal variances with naive penalty (top left panel), longitude-wise circular penalty (top right panel) and geodesic penalty (bottom panel). (The colour bar on the right is valid for all the three panels).

prior for the noise precision is negligible. In general, choice about the prior $\pi(\tau_\beta)$ will be relevant in situations where we have a poor sample size compared to the number of parameters which require to be estimated. In the case study under examination the large sample size available for estimating each spline coefficient makes the impact of $\pi(\tau_\beta)$ negligible.

#### *3.2.5. Computations*

Model estimation does not raise particular issues with respect to planar P-spline models, once matrices $\boldsymbol{B}$ and $\boldsymbol{R}$ have been built. Indeed, the model belongs to the class of Latent Gaussian Markov Models and approximate Bayesian inference can be performed efficiently using the `R-INLA` package (Rue et al., 2009). In our case study, we find it more appropriate to use a Gibbs sampling algorithm as the tools developed for detecting the ITCZ (see section 4.3) require a sample from the joint posterior distribution of the model. The most expensive step is to sample from the full conditional for the spline coefficients

$$\boldsymbol{\beta}|\tau_\beta, \tau_\epsilon, \boldsymbol{y} \sim N(\boldsymbol{Q}^{-1}\boldsymbol{B}^\top \boldsymbol{y}, \boldsymbol{Q}^{-1}) \qquad \boldsymbol{Q} = \left(\boldsymbol{B}^\top \boldsymbol{B} + \frac{\tau_\beta}{\tau_\epsilon}\boldsymbol{R}\right) \tag{5}$$

under the linear constraint $\mathbf{1}_K^\top \boldsymbol{B}\boldsymbol{\beta} = 0$ needed for intercept identifiability. We use an efficient Gibbs sampler coded in `R` with the use of sparse matrix algebra as implemented in the `spam` package (Furrer and Sain, 2010) to exploit sparsity of $\boldsymbol{Q}$ in (5). The `spam` package contains routines to perform an efficient Cholesky decomposition of $\boldsymbol{Q}$, which is important for fast sampling from a GMRF under linear constraints like the full conditional $\pi(\boldsymbol{\beta}|\tau_\beta, \tau_\epsilon, \boldsymbol{y})$ in (5). The full conditionals for all the parameters in the model and the code for implementing the Gibbs sampler in `R` can be found in the supplementary material.

## 4. Application

### *4.1. Modelling TCWV data*

The goal of our application is to detect the ITCZ location by using the TCWV dataset described in Section 2. The operative definition of ITCZ that we use, as suggested by researchers from ISAC-CNR, Italy, is "the strip surrounding the Earth surface where TCWV shows highest values".

To this aim, we first apply Geodesic P-splines for smoothing observed TCWV data, which is affected by noise and does not provide measurements over the land, in order to predict the latent field all over the world. Then, we exploit the model output for locating ITCZ by sampling from the joint posterior distribution of the latent field. A set of $m = 1000$ data randomly scattered over the Earth's surface is held out from model estimation for validation purposes (see section 4.2). We denote with $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ the vector of TCWV observations used for model estimation and with $\boldsymbol{y}^* = (y_1^*, \ldots, y_m^*)^\top$ the vector of validation data. We fitted the Geodesic P-spline model to the data displayed in Figure 1, referred to January and July, 2008.

The hierarchical model is

$$
\begin{array}{rcll}
\text{Likelihood:} & & & \\
\boldsymbol{y}|\alpha, \boldsymbol{\beta}, \tau_\epsilon & \sim & \mathcal{N}(\boldsymbol{\mu}, \tau_\epsilon^{-1}\boldsymbol{I}) & \\
\boldsymbol{\mu} & = & \alpha\mathbf{1} + \boldsymbol{B}\boldsymbol{\beta} & \qquad (6) \\
\text{Prior:} & & & \\
\alpha & \sim & \mathcal{N}(0, \tau_\alpha^{-1}) & \\
\boldsymbol{\beta}|\tau_\beta & \sim & \mathcal{N}(\mathbf{0}, \tau_\beta^{-1}\boldsymbol{R}^*) \qquad \boldsymbol{\beta} \text{ is subject to } \mathbf{1}_K^\top \boldsymbol{B}\boldsymbol{\beta} = 0 & \qquad (7) \\
\text{Hyper-prior:} & & & \\
\tau_\beta & \sim & \mathrm{Ga}(1, 5e-5) & \\
\tau_\epsilon & \sim & \mathrm{Ga}(1, 5e-5) &
\end{array}
$$

At the likelihood level, the matrix $\boldsymbol{B}$ in (6) is the B-spline basis on a GDGG as described in Section 3.2. The latent field $\boldsymbol{\mu}$ is a surface varying smoothly over the sphere, with $\alpha$ the global spatial mean and $\boldsymbol{\beta}$ the spline coefficients. At the prior level, we have a diffuse Gaussian prior, with $\tau_\alpha$ fixed at a small value for the intercept and an ICAR prior, with precision $\tau_\beta\boldsymbol{R}^*$, for the spline coefficients. Using the scaled matrix $\boldsymbol{R}^*$ is a fundamental step: this allows us to select the same prior for $\tau_\beta$ and $\tau_\epsilon$, as both $\boldsymbol{I}$ in (6) and $\boldsymbol{R}^*$ in (7) have average marginal variance equal to 1. The results presented in this section are obtained using a $\mathrm{Ga}(1, 5e-5)$ for both hyperparameters, after checking that the results were non sensitive to other choices for $a$ and $b$.

To compute $\boldsymbol{B}$, the latitude and longitude coordinates are converted into spherical coordinates, then projected on the icomesh and finally cubic B-splines are evaluated on a GDGG with $K$ knots, where $K$ depends on $\nu$, the number of split iterations performaed on the icosahedron: the choice of $\nu$ is a critical aspect of the method and will be discussed in section 4.3, where we compare results obtained with $\nu = 1, \ldots, 6$.

Model estimation is performed by Gibbs sampling: we draw a total of 5000 samples after convergence (achieved after a quick burnin due to the large sample size available for each model parameter). Regarding computational time, it takes about ten seconds (in a laptop with Intel core i7, 2.5GHz, 16 GB ram) to run a hundred iterations when $\nu = 5$.

In a Bayesian framework, spatial prediction is naturally based on the joint posterior predictive distribution of the latent field: sampling from this distribution is particularly efficient when using Bayesian P-splines. Once the posterior distribution of the latent field $\pi(\boldsymbol{\mu}|\boldsymbol{y})$ has been obtained, prediction $\tilde{\mu}$ at an arbitrary location $\tilde{\boldsymbol{x}}$ can be performed, after evaluating the basis functions at $\tilde{\boldsymbol{x}}$, using the posterior predictive distribution:

$$\pi(\tilde{\mu}|\boldsymbol{y}) = \int \pi(\tilde{\mu}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \tag{8}$$

where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}, \tau_\alpha, \tau_\beta, \tau_\epsilon)$. This is achieved by composite sampling once $G$ samples from the posterior distribution are available. Samples from distribution (8) are obtained by sampling from $\pi(\tilde{\mu}|\boldsymbol{\theta}^g)$, where $\boldsymbol{\theta}^g$ is an MCMC sample from the posterior distribution of $\boldsymbol{\theta}$.

### *4.2. Model checking*

In this section the goal is twofold: firstly, we investigate the predictive performance of the proposed Geodesic P-spline (G-Pspline) model for different choices of $\nu$. Secondly, we compare the G-Pspline and the stochastic partial differential equation (SPDE) approach by Lindgren et al. (2011), both in terms of computational and predictive perfomance, using the TCWV data.

The predictive performance is evaluated by first estimating the model on training dataset and then computing error measures on a validation dataset. Let $\hat{y}^*_j = E(y^*_j|\boldsymbol{y})$ denote the mean of the posterior predictive distribution at the validation location $j$, $j = 1, \ldots, m$, as measures of predictive performances we consider both the relative mean absolute prediction error (RMAPE),

$$RMAPE = \frac{1}{m}\sum_{j=1}^{m}\left|\frac{\hat{y}^*_j - y^*_j}{y^*_j}\right|, \tag{9}$$

where the average is taken over the validation locations, and the relative mean square prediction error (RMSPE), which is the same as (9) except for averaging squares, instead of absolute, relative errors.

#### *4.2.1. G-Pspline model performance for varying $\nu$*

The first three lines of Table 2 report the RMAPE and the RMSPE for the G-Pspline model for different number of knots, $K$ (note that the different $K$'s are associated to

$\nu = 1, \ldots, 6$). The RMSPE shows a minimum at $\nu = 5$ (0.109) while the RMAPE is almost unchanged when increasing $\nu$ from 5 to 6 as it decays from 12.8% ($\nu = 5$) to 12.2% ($\nu = 6$). Based on these results we find appropriate to select $\nu = 5$ for ITCZ location, as this allows us to gain computational speed in the procedure described in section 4.3.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **G-Pspline** | $K$ | 42 | 162 | 642 | 2562 | 10242 | 40962 |
| | RMAPE | 0.263 | 0.213 | 0.169 | 0.146 | 0.128 | 0.122 |
| | RMSPE | 0.218 | 0.206 | 0.166 | 0.136 | 0.109 | 0.122 |
| **SPDE** | $K$ | 43 | 164 | 644 | 2580 | 10243 | 40841 |
| | RMAPE | 0.518 | 0.204 | 0.166 | 0.148 | 0.128 | 0.118 |
| | RMSPE | 1.619 | 0.186 | 0.163 | 0.151 | 0.106 | 0.110 |

Table 2: Relative mean absolute prediction error (RMAPE) and relative mean square prediction error (RMSPE) obtained with Geodesic P-Splines (G-Pspline) and SPDE. $K$ denotes the number of knots of the geodesic grid for G-Pspline (for $\nu = 1, \ldots, 6$), or the number of knots of the triangular mesh for SPDE (obtained by tuning the `max.edge` argument in the `R-INLA` function `inla.mesh.2d`).

The choice of $\nu$ is the starting step of the modelling process, similarly to the choice of a suitable triangulation in the SPDE approach. Literature on P-splines recommends using a number of knots $K$ large enough to describe the spatial variation of the data and let the penalty determine the right amount of smoothing. Under different $K$ levels, provided that $K$ is large enough, the same degree of smoothing is obtained by rescaling the smoothing parameter accordingly. This is confirmed in Table 2, where measures of predictive performance of the G-Pspline model remain almost unchanged for $K = 10242$ ($\nu = 5$) and $K = 40962$ ($\nu = 6$). In a Bayesian P-spline setting, this rescaling is reflected in the posterior distribution of $\tau_\beta$; if $K$ changes, the location of $\pi(\tau_\beta|\boldsymbol{y})$ shifts accordingly.

### 4.2.2. *Comparing G-Pspline and SPDE*

SPDE is implemented in the `R-INLA` package (Martins et al., 2013) and, as a starting point, needs the definition of a triangular mesh covering the study region, analogously to the definition of a geodesic grid in our framework. For the sake of comparison of the prediction performance the SPDE mesh and the G-Psline geodesic grid should have similar size. For each column in Table 2 (i.e., for each $\nu$ from 1 to 6) the triangular mesh is built using the `R-INLA` function `inla.mesh.2d`, by tuning the `max.edge` argument (the largest allowed triangle edge length) in order to have roughly the same number of knots $K$.

Looking at Table 2 column-wise, G-Pspline and SPDE perform similarly in terms of RMAPE and RMSPE. The boxplots in Figure 6 show the variability of the (log)

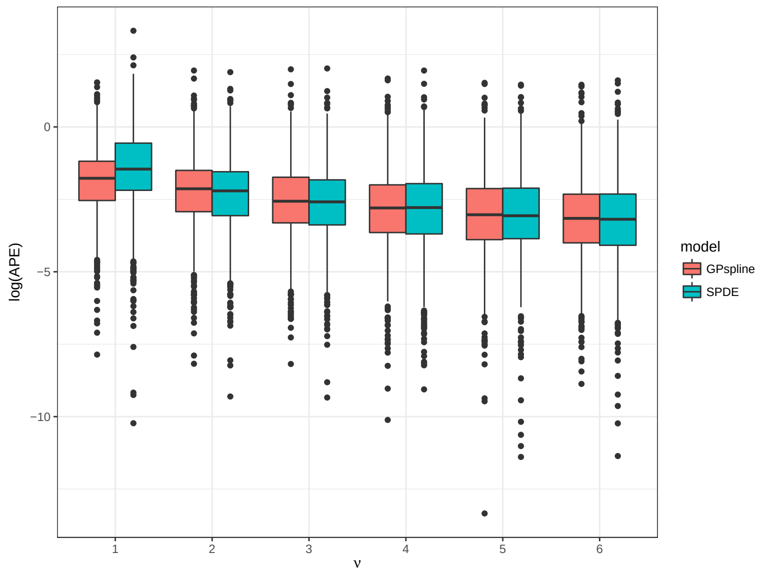


Figure 6: Boxplots of the (log) relative absolute prediction errors (APE), $\log(|\hat{y}^*_j - y^*_j/y^*_j|)$, measured on a validation set of $m = 1000$ locations. Prediction performance for the G-Pspline and the SPDE models is very similar for each $\nu$.

relative absolute prediction errors over the $m$ validation locations, for both models and varying $\nu$. For each $\nu$, the variability of the relative absolute prediction errors is practically the same for G-Pspline and SPDE. Boxplots for the squared prediction errors present a similar pattern and are not shown here. Based on these results we can conclude that, in our case study, prediction performance measured on a validation set of $m = 1000$ locations is overall very similar for G-Pspline and SPDE.

As a final note on computation time, the SPDE model fitted within `R-INLA` is faster than the G-Pspline fitted via Gibbs sampling: SPDE takes around four minutes, while our Gibbs sampler takes around ten minutes to run 5000 iterations. Nonetheless, the procedure described in Section 4.3 for ITCZ location requires MCMC samples from the model posterior. Sampling from the posterior of the latent field within `R-INLA` (using `inla.posterior.sample()`) is computationally intensive for our model, as it takes around 30 seconds to run 10 samples. Therefore, to the purpose of ITCZ location, the proposed Geodesic P-spline approach using Gibbs sampling is overall faster than SPDE within `R-INLA`, while maintaning the same predictive performance. Results on the ITCZ location obtained with SPDE for January and July (not shown here) were very similar to those presented in Figure 8 which are obtained using the procedure discussed next.

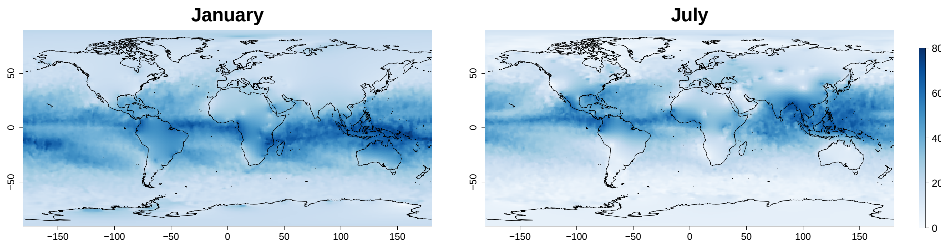


Figure 7: Model prediction of the latent field for January and July.

### 4.3. Locating the ITCZ

In Figure 7 we illustrate the maps of the TCWV posterior means obtained with $\nu = 5$: this accomplishes our first task, i.e. to remove random noise from data and to reconstruct the latent field on the whole of Earth's surface.

The problem of ITCZ location is addressed by summarising the posterior predictive distribution of the TCWV latent field. The procedure outlined below requires the specification of a reasonable guess concerning the width of the ITCZ region denoted as $W$; we based our choice on expert knowledge by ISAC-CNR researchers and set $W = 1000$ *km*. The ITCZ width relative to the length of a Meridian (which is about 20000 *km*) is around $w = W/20,000 = 0,05$.

Our algorithm to locate the ITCZ consists of a discrete search performed longitude-wise (i.e. at each meridian). Let $m = 1, \ldots, M$ indicate a set of $M$ meridians: for a given $m$, we sample from the posterior predictive distribution of the latent field at a fine grid over latitude. Then, we compute the posterior probability that a point at a given latitude belongs to the region where the TCWV shows the highest values (i.e. the point falls into the ITCZ region), integrating out uncertainty about model parameters.

Let $\tilde{\boldsymbol{\mu}}_m = (\tilde{\mu}_{1m}, \ldots, \tilde{\mu}_{lm}, \ldots, \tilde{\mu}_{Lm})$ be the vector of the latent field predicted at locations $l = 1, \ldots, L$, where $(lat_{1m}, \ldots, lat_{lm}, \ldots, lat_{Lm})$ is a regular sequence from 90° to −90°. The algorithm proceeds as follows. For $m = 1, \ldots, M$:

- evaluate the bases at locations $l = 1, \ldots, L$, this gives a meridian-specific $L \times K$ dimensional basis matrix $\tilde{\boldsymbol{B}}_{\boldsymbol{m}}$;
- sample $G$ realizations from the posterior predictive distribution (8) by computing $\tilde{\boldsymbol{\mu}}_m^g = \alpha^g \mathbf{1} + \tilde{\boldsymbol{B}}_{\boldsymbol{m}} \hat{\boldsymbol{\beta}}^g$, $g = 1, \ldots, G$;
- for $g = 1, \ldots, G$, rank the vector $\tilde{\boldsymbol{\mu}}_m^g$. This gives a posterior sample of the ranks, indicated by vector $\boldsymbol{\phi}_m^g = (\phi_{1m}^g, \ldots, \phi_{lm}^g, \ldots, \phi_{Lm}^g)$, e.g. $\phi_{lm}^g = L$ if $l = \text{argmax}_l(\tilde{\boldsymbol{\mu}}_m^g)$, while $\phi_{lm}^g = 1$ if $l = \text{argmin}_l(\tilde{\boldsymbol{\mu}}_m^g)$.

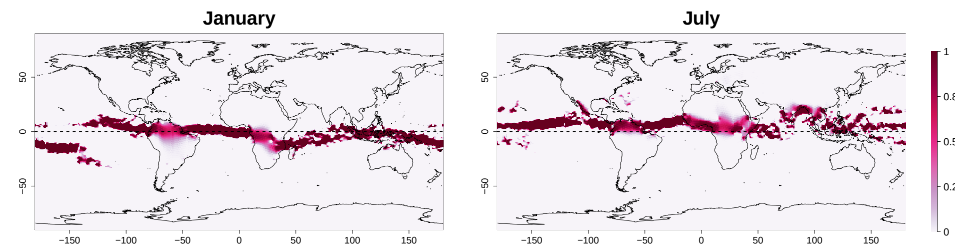


Figure 8: ITCZ location for January and July.

The probability that a point $l$ belonging to meridian $m$ falls into the ITCZ is computed as

$$Pr\left(lat_{lm} \in ITCZ|\mathbf{y}\right) = \frac{1}{G}\sum_{g=1}^{G} \mathbb{1}\left(1 - \frac{\phi_{lm}^{g}}{L} < w\right) \tag{10}$$

where $\mathbb{1}$ is the indicator function and $\phi_{lm}^{g}/L$ is the normalised rank. To sum up the above, (10) is the probability that the point with geographical coordinates $(lat_l, long_m)$ falls inside the ITCZ, where the length of the ITCZ is fixed according to $W$. Results are displayed in Figure 8 for the two months under examination: this Figure is obtained running the algorithm with $L = 1000$ and $M = 360$. The ITCZ is mostly located in the south (north) of the Equator in January (July), as expected on the basis of prior knowledge concerning its seasonal behaviour. The map for January shows the double ITCZ, which is typical of the Central Pacific region in some period during the year (Waliser and Gautier, 1993). The proposed method allows to locate the ITCZ even over land (in particular in Africa and South America) where data is not available, this being reflected by higher posterior uncertainty. Of course, the width of ITCZ reported in Figure 8 is strictly dependent on the choice of $W$: although this is very relevant when studying a single month, we believe that it is not such a crucial choice if the method is used for studying the spatio-temporal trend of the phenomenon. Indeed, in this case it would be important to keep $W$ fixed along the study period in order to ensure comparability among results.

## 5. Discussion

We presented a Bayesian hierarchical framework for smoothing data collected worldwide at a large number of locations. With respect to traditional methods, the proposed model accounts for geodesic distances between the data, thus overcoming the limitations of covariance functions for Euclidean spaces when applied to global datasets.

The non-parametric model formulation proposed extends the Bayesian P-spline approach for smoothing worldwide collected data. Using a sphere as a representation of the Globe, the idea is to build a new basis of B-splines on a suitable geodesic grid while keeping the hierarchical model formulation of Bayesian P-splines, with the associated advantages in terms of flexibility and computation. Two key features of P-splines are maintained in the Geodesic P-spline model: (a) the use of local bell-shaped functions, e.g. the B-splines on the icomesh, that yield a sparse basis matrix; (b) the use of B-splines centred at equidistant knots, i.e. the nodes of the icomesh. Point (b) suggests that an IGMRF for regular locations is a sensible prior distribution for the spline coefficients, giving stable marginal variances as opposed to the standard P-spline model construction. Computational efficiency is due to (a) reduction of the latent field dimension, as the smoothing prior operates on the spline coefficients (low-rank smoother) and (b) fast MCMC based on sparse Cholesky factorization of the structure matrix of the full conditional for the latent field. These advantages allow for a fast fitting of the model to data collected worldwide in high-resolution.

We applied the Geodesic P-spline model to the TCWV data retrieved with the AIRWAVE algorithm at a huge number of locations on Earth. The smoothing approach in this example is desirable as it allows field estimation at unmonitored locations. We provided inferential tools to locate the ITCZ based on ranking samples from the posterior distribution of the latent field, estimated at a fine grid over the Globe. Results are coherent with prior knowledge concerning ITCZ, indicating a shift towards southern regions in autumn and winter.

A critical aspect is the choice of hyperprior for the random walk precision, $\pi(\tau_\beta)$. We expect a large impact of $\pi(\tau_\beta)$ in situations where sample size is small compared to the number of parameters required to be estimated. In the case study on TCWV, the sample size available for estimating each spline coefficients is large enough, which makes the impact of $\pi(\tau_\beta)$ very small. In the results presented in Section 4 we used a Gamma with shape $a = 1$ and rate $b = 5e-5$ for both $\tau_\beta$ and $\tau_\epsilon$, after checking that the posterior $\pi(\tau_\beta|\boldsymbol{y})$ remained unchanged under different choices of $a$ and $b$. We believe that controlling that the posterior learns from the data in the same way for different choices of the prior is a reasonable approach to test the robustness of the Bayesian specification. On the topic of prior selection for variance parameters the literature has shown rapid growth over the past decade; see, e.g., Gelman (2006); Simpson et al. (2017) and references therein.

The model can be extended in several directions, both in a methodological and applied sense. In this paper we focused on an IGMRF structure for the spline coefficients equivalent to the ICAR used for lattice data, using the six surrounding knots as neighbours. Investigation of geodesic grids suitable for higher order IGMRF priors would be interesting. Another attractive research line would be to look into a model based on nested B-splines, defined on a set of geodesic grids of different resolution, following Nychka et al. (2015). In a fully Bayesian framework this requires careful hyperprior specification, as it is not clear how to prevent confounding between nested components.

As for application, a future research line worthy of investigation is modelling the ITCZ based on different proxy variables, focusing the analysis on a wide temporal range, following the ideas in Castelli et al. (2017). The application of Geodesic P-spline models to the 20 years of ATSR data will allow the investigation of ITCZ meridional migration trends. Moreover, joint modelling of TCWV and other ITCZ related phenomena, possibly available at misaligned locations, will result in more reliable estimates of the ITCZ latent field, especially at locations where TCWV retrieval is not possible with current ATSR technology.


## Acknowledgements

The work by Fedele Greco and Massimo Ventrucci is funded by the PRIN2015-supported project *Environmental processes and human activities: capturing their interactions via statistical methods* (EPHASTAT) by MIUR (Italian Ministry of Education, University and Scientific Research). We thank Bianca Maria Dinelli (ISAC-CNR), Enzo Papandrea and Stefano Casadio (Serco s.p.a.) for guidance in the interpretation of the results. ATSR TCWV dataset was developed in the frame of the ESA ALTS project ESA Contract No. 4000108531/13/I-NB.


## References


Banerjee, S., 2005. On Geodetic Distance Computations in Spatial Modeling. Biometrics 61 (2), 617–625.
URL `http://dx.doi.org/10.1111/j.1541-0420.2005.00320.x`

Banerjee, S., 2017. High-dimensional bayesian geostatistics. Bayesian Analysis 12 (2), 583–614.
URL `https://doi.org/10.1214/17-BA1056R`

Banerjee, S., Gelfand, A., Carlin, B., 2014. Hierarchical Modeling and Analysis for Spatial Data, Second Edition. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Banerjee, S., Gelfand, A. E., Finley, A. O., Sang, H., 2008. Gaussian predictive process models for large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (4), 825–848.
URL `http://dx.doi.org/10.1111/j.1467-9868.2008.00663.x`

Baramidze, V., Lai, M. J., Shum, C. K., 2006. Spherical splines for data interpolation and fitting. SIAM Journal on Scientific Computing 28 (1), 241–259.
URL `https://doi.org/10.1137/040620722`

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology) 36(2), 192–225.

Casadio, S., Castelli, E., Papandrea, E., Dinelli, B., Pisacane, G., Burini, A., Bojkov, B., 2016. Total column water vapour from along track scanning radiometer series using thermal infrared dual view ocean cloud free measurements: The advanced infra-red water vapour estimator (airwave) algorithm. Remote Sensing of Environment 172, 1–14.

Castelli, E., Ventrucci, M., Greco, F., Valeri, M., Dinelli, B., Papandrea, E., Casadio, S., 2017. On the contribution of 20 years of ATSR data and geodesic P-spline: Efficient spatial smoothing method to ITCZ trend analysis. In: Proceedings of the 2017 conference on Big Data from Space (BiDS17), 28-30 October 2017, Toulouse, France.

Cressie, N., Johannesson, G., 2008. Fixed rank kriging for very large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (1), 209–226.
URL `http://dx.doi.org/10.1111/j.1467-9868.2007.00633.x`

Datta, A., Banerjee, S., Finley, A. O., Gelfand, A. E., 2016. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. Journal of the American Statistical Association 111 (514), 800–812.
URL `https://doi.org/10.1080/01621459.2015.1044091`

Di Marzio, M., Panzera, A., Taylor, C. C., 2014. Nonparametric regression for spherical data. Journal of the American Statistical Association 109 (506), 748–763.
URL `http://dx.doi.org/10.1080/01621459.2013.866567`

Duchamp, T., Stuetzle, W., 2003. Spline smoothing on surfaces. Journal of Computational and Graphical Statistics 12 (2), 354–381.
URL https://doi.org/10.1198/1061860031743

Eilers, P., Currie, I., Durbán, M., 2006. Fast and compact smoothing on large multidimensional grids. Computational Statistics & Data Analysis 5, 61–76.

Eilers, P., Marx, B., 1996. Flexible Smoothing with B-splines and Penalties. Statistical Science 11, 89–121.

Ettinger, B., Perotto, S., Sangalli, L. M., 2016. Spatial regression models over two-dimensional manifolds. Biometrika 103 (1), 71–88.
URL http://dx.doi.org/10.1093/biomet/asv069

Furrer, R., Genton, M. G., Nychka, D., 2006. Covariance tapering for interpolation of large spatial datasets. Journal of Computational and Graphical Statistics 15, 502–523.

Furrer, R., Sain, S., 2010. spam: A Sparse Matrix R Package with Emphasis on MCMC Methods for Gaussian Markov Random Fields. Journal of Statistical Software, Articles 36 (10), 1–25.
URL https://www.jstatsoft.org/v036/i10

Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis 1 (3), 515–534.
URL http://dx.doi.org/10.1214/06-BA117A

Gneiting, T., 2013. Strictly and non-strictly positive definite functions on spheres. Bernoulli 19 (4), 1327–1349.
URL http://dx.doi.org/10.3150/12-BEJSP06

Holton, J. R., Hakim, G. J., 2013. An Introduction to Dynamic Meteorology (Fifth Edition). Academic Press, Boston.
URL https://www.sciencedirect.com/science/article/pii/B9780123848666000301

Jeong, J., Jun, M., 2015. A class of matérn-like covariance functions for smooth processes on a sphere. Spatial Statistics 11, 1 – 18.
URL http://www.sciencedirect.com/science/article/pii/S2211675314000554

Jun, M., Stein, M. L., 2007. An approach to producing space: Time covariance functions on spheres. Technometrics 49 (4), 468–479.
URL http://www.jstor.org/stable/25471392

Lai, M., Schumaker, L., 2007. Spline Functions on Triangulations. No. v. 13 in Encyclopedia of Mathematics an. Cambridge University Press.
URL https://books.google.it/books?id=6hvqGgbBmEoC

Lai, M. J., Shum, C. K., Baramidze, V., Wenston, P., Aug 2009. Triangulated spherical splines for geopotential reconstruction. Journal of Geodesy 83 (8), 695–708.
URL https://doi.org/10.1007/s00190-008-0283-0

Lang, S., Brezger, A., 2004. Bayesian p-splines. Journal of Computational and Graphical Statistics 13, 183–212.

Lee, D., 2010. Smothing mixed model for spatial and spatio-temporal data. Tech. rep., Ph.D. Thesis, Department of Statistics, Universidad Carlos III de Madrid, Spain.

Lee, D., Durbán, M., 2009. Smooth-car mixed models for spatial count data. Computational Statistics and data Analysis 53, 2968–2977.

Lindgren, F., Rue, H., Lindstrm, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73 (4), 423–498.
URL http://dx.doi.org/10.1111/j.1467-9868.2011.00777.x

Martins, T. G., Simpson, D., Lindgren, F., Rue, H., 2013. Bayesian computing with inla: New features. Computational Statistics & Data Analysis 67 (0), 68 – 83.
URL http://www.sciencedirect.com/science/article/pii/S0167947313001552

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., Sain, S., 2015. A multiresolution gaussian process model for the analysis of large spatial datasets. Journal of Computational and Graphical Statistics 24 (2), 579–599.
URL http://dx.doi.org/10.1080/10618600.2014.914946

Porcu, E., Bevilacqua, M., Genton, M. G., 2016. Spatio-temporal covariance and cross-covariance functions of the great circle distance on a sphere. Journal of the American Statistical Association 111 (514), 888–898.
URL https://doi.org/10.1080/01621459.2015.1072541

Randall, D. A., Ringler, T. D., Heikes, R., Jones, P., Baumgardner, J., 2002. Climate modeling with spherical geodesic grids. Computing in Science and Engineering 4, 32–41.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields. Chapman and Hall/CRC.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 (2), 319–392.
URL http://dx.doi.org/10.1111/j.1467-9868.2008.00700.x

Sahr, K., White, D., Kimerling, A. J., 2003. Geodesic Discrete Global Grid Systems. Cartography and Geographic Information Science 30 (2), 121–134.
URL http://dx.doi.org/10.1559/152304003100011090

Sangalli, L. M., Ramsay, J. O., Ramsay, T. O., 2013. Spatial spline regression models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75 (4), 681–703.
URL http://dx.doi.org/10.1111/rssb.12009

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Srbye, S. H., 02 2017. Penalising model component complexity: A principled, practical approach to constructing priors. Statist. Sci. 32 (1), 1–28.
URL https://doi.org/10.1214/16-STS576

Sørbye, S., Rue, H., 2014. Scaling intrinsic gaussian markov random field priors in spatial modelling. Spatial Statistics 8, 39 – 51.
URL http://www.sciencedirect.com/science/article/pii/S2211675313000407

Ugarte, M., Goicoa, T., Etxeberria, J., Militino, A., 2012. A p-spline anova type model in space-time disease mapping. Stochastic Environmental Research and Risk Assessment 26 (6), 835–845.

Wahba, G., 1981. Spline interpolation and smoothing on the sphere. SIAM Journal on Scientific and Statistical Computing 2 (1), 5–16.
URL https://doi.org/10.1137/0902002

Waliser, D. E., Gautier, C., 1993. A satellite-derived climatology of the ITCZ. Journal of Climate 6 (11), 2162–2174.

Wikle, C., Cressie, N., 1999. A dimension-reduced approach to space-time Kalman filtering. Biometrika 86 (4), 815–829.
URL +http://dx.doi.org/10.1093/biomet/86.4.815

Wilhelm, M., Ded, L., Sangalli, L. M., Wilhelm, P., 2016. Igs: An isogeometric approach for smoothing on surfaces. Computer Methods in Applied Mechanics and Engineering 302, 70 – 89.
URL http://www.sciencedirect.com/science/article/pii/S0045782516000025

Wood, S., 2017. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.

Žagar, N., Skok, G., Tribbia, J., 2011. Climatology of the ITCZ derived from ERA Interim reanalyses. Journal of Geophysical Research: Atmospheres 116 (D15).

Zhang, C., 2001. Double ITCZs. Journal of Geophysical Research: Atmospheres 106 (D11), 11785–11792.

Heaton, M., Katzfuss, M., Berrett, C., Nychka, D., 2014. Constructing valid spatial processes on the sphere using kernel convolutions. Environmetrics 25 (1), 2–15.
URL http://dx.doi.org/10.1002/env.2251