Multi-objective environmental model evaluation by means of multidimensional kernel density estimators: efficient and multi-core implementations

Unai Lopez-Novoa[1], Jon Sáenz[2,3], Alexander Mendiburu[1], Jose Miguel-Alonso[1], Iñigo Errasti[4], Ganix Esnaola[2], Agustín Ezcurra[2], Gabriel Ibarra-Berastegi[4]

Corresponding author: Unai Lopez-Novoa {E-mail: unai.lopez@ehu.es, Phone: +34 943 018 012}

[1]Intelligent Systems Group, Dept. of Computer Architecture and Technology, University of the Basque Country (UPV/EHU), P. Manuel Lardizabal 1, 20018, Donostia-San Sebastian, Spain [2]Dept. Applied Physics II, University of the Basque Country (UPV/EHU), Sarriena Auzoa z/g, 48940-Leioa, Spain.

[3]Plentzia Marine Station (PIE-UPV/EHU), Areatza Pasealekua, 48620, Plentzia, Spain

[4]Dept. Nuclear Engineering and Fluid Mechanics, University of the Basque Country (UPV/EHU), Alda. Urkijo, s/n, 48013, Bilbao, Spain

Abstract

We propose an extension to multiple dimensions of the univariate index of agreement between PDFs used in climate studies. We also provide a set of high-performance programs targeted both to single and multi-core processors. They compute multivariate PDFs by means of kernels, the optimal bandwidth using smoothed bootstrap and the index of agreement between multidimensional PDFs. Their use is illustrated with two case-studies. The first one assesses the ability of seven global climate models to reproduce the seasonal cycle of zonally averaged temperature. The second case study analyzes the ability of an oceanic reanalysis to reproduce global sea surface temperature and sea surface height. Results show that the proposed methodology is robust to variations in the optimal bandwidth used. The technique is able to process multivariate datasets corresponding to different physical dimensions. The methodology is very sensitive to the existence of a bias in the model with respect to observations.

Keywords: Multivariate kernel density estimation**,** multidimensional kernel density estimation, multi-core implementation, environmental model evaluation

This is the accepted manuscript of the article that appeared in final form in **Environmental Modelling & Software** 63 : 123-136 (2015), which has been published in final form at https://doi.org/10.1016/j.envsoft.2014.09.019.

Software availability:

Name or software or dataset: density-parallel

Developers: Unai Lopez-Novoa (1) Jon Sáenz (2) Alexander Mendiburu (1) Jose Miguel-Alonso (1)

(1) Intelligent Systems Group, Dept. of Computer Architecture and Technology, University of the Basque Country (UPV/EHU), P. Manuel Lardizabal 1, 20018, Donostia-San Sebastian, Spain. unai.lopez@ehu.es, Phone: +34 943 018 012 , alexander.mendiburu@ehu.es, Phone: +34 943015020 and j.miguel@ehu.es Phone: +34 943 018019

(2) EOLO Group, Department of Applied Physics II, University of the Basque Country, (UPV/EHU), Barrio Sarriena s/n, 48940-Leioa, Spain. jon.saenz@ehu.es, Phone: +34 946012445.

Year first available: 2014

Hardware required: any single-core or multi-core processor.

Software required:

- C Compiler (e.g. GCC). OpenMP compiling capabilities required for multi-threaded implementations
- MESCHACH math library. Available at http://homepage.math.uiowa.edu/~dstewart/meschach/
- netCDF data handling library. Available at http://www.unidata.ucar.edu/software/netcdf/

Program language: ANSI-C

Program size, including example data: 102 KB in a single .tar.gz file.

License: Open software, available under a “New BSD 3-clause license”.

Availability: http://www.sc.ehu.es/ccwbayes/isg/index.php?option=com_remository&Itemid=13 Or go to “Downloads” section of http://www.sc.ehu.es/isg

1 Introduction

Climate models are the best tools that scientists currently have in order to assess the impact of increasing concentrations of greenhouse gases and other anthropogenic influences on the observed climate of the Earth. These tools allow scientists to understand climatic changes from a dynamical point of view and to give quantitative answers to questions about future climate. They make feasible the assessment of the characteristics (deterministic versus stochastic) of some climatic variations at different temporal or spatial scales. Finally, they are fundamental in the attribution phase of the study of the climate change problem, since they allow to confidently discard competing hypothesis such as whether the climate change is rooted in natural or anthropogenic causes (Bengtsson, 2013; Knutti, 2008).

Climate models must be evaluated against different observations (Otto et al., 2013) or paleoclimate data (Braconnot et al., 2012; Hind et al., 2012; Moberg, 2013; Sundberg et al., 2012) in order to get a quantitative indication on the confidence that we can put into their outputs depending on the efficacy of the models to reliably represent the climatic processes and feedbacks. Contrary to operational weather forecast models, there is no way to properly evaluate models against future climate, since future climate does not exist yet (Randall et al., 2007; Stocker et al., 2013). Additionally, since parameterizations of sub-grid scale processes are not fully independent from current climate, it is clear that evaluation against current climate is the only feasible, albeit not perfect, solution in terms of evaluating the projections for future climate (Errasti et al., 2011, 2013; Radić and Clarke, 2011; Reichler and Kim, 2008).

The rationale behind this hypothesis is that models that are able to better simulate current climate are the ones that we expect will also be the best ones in terms of the simulation of future climate. It is well known that this is not necessarily true due to the different behavior of models in terms of their internal feedback mechanisms (Andrews et al., 2012; Dessler, 2013). These feedbacks lead to differing values of the climate sensitivity of models and, hence, future warming is dependent on these different sensitivities, leading to the question whether all the models are equally valid (Knutti, 2010).

Particularly for downscaling applications and regional impact analysis, an evaluation of the adequacy of models is a common step (Brands et al., 2011; Radić and Clarcke, 2011; Walsh et al., 2008). Considering that numerical downscaling is computationally expensive, it cannot be

performed over all the models available from a big experiment such as CMIP3 or CMIP5, and it is usually only performed on a subset of the models (Hewitt and Griggs, 2004). Then, the best models are only used for downscaling in regional climate assessment, since they have been proven to be the most suitable over a particular region. This strategy of selecting the best subset of all the available models has been contested by some studies (Reifen and Toumi, 2009) and defended by others (Macadam et al., 2010), since this result depends on the removal of the seasonal cycle and the use of anomalies instead of raw output from the models.

There are several inter-comparison experiments that have been set-up in order to drive the models with common boundary conditions so that results between model runs can be compared. As an example of this kind of standard experimental setups, that in several cases have their origins in the nineties, we can cite the Atmospheric Model Intercomparison Project (Gates et al., 1998), the Project for Intercomparison of Land Surface Parameterization Schemes (PILPS) (Henderson-Sellers et al., 1995) or the Palaeoclimate Modelling Intercomparison Project (PMIP) (Kageyama et al., 1999), the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP), described by Lamarque et al. (2013), or the one that is most relevant for our study, since we use data from this experiment, the Coupled Model Intercomparison Project (CMIP) (Meehl et al., 2007; Taylor et al., 2012). These projects have covered several phases through the years and for the case of the CMIP data, CMIP5 can already be used. In general, models grouped under a similar experimental set-up such as CMIP3 or CMIP5 are considered as an ensemble of opportunity (Annan and Hargreaves, 2010). There are some limitations because even for coordinated experiments such as CMIP3, there is some freedom in the way the external boundary conditions are applied (they are not 100% equal for all the models), see Table 1 in Wang et al. (2007). The number of realizations from every model in the ensemble of opportunity is not the same, neither and, therefore, the influence of each model in the behaviour of the ensemble is not the same. Additionally, models are less independent than they should be, since critical algorithms or components are shared by several models (Fernández et al., 2009; Knutti et al., 2013; Masson and Knutti, 2011; Pennell and Reichler, 2011).

In terms of evaluation of climate models, it is well known that climate simulations are run, most of the times, past the limit of deterministic predictability associated with predictability of the first kind, according to Lorenz's classification. Climate models simulate climate change under varying boundary conditions in terms of the Probability Density Functions (PDF) of climatic variables. The varying boundary conditions consist of external forcings such as the variability in solar irradiance,

orbital parameters or anthropogenic greenhouse gas emissions, amongst other potential driving factors (Bengtsson, 2013). Some aspects of climate simulations are deterministic, such as the seasonal cycle at extratropical latitudes (Errasti et al., 2013). On the other side, some properties of the atmospheric circulation such as the blocking at extratropical latitudes can not be precisely forecast with lead times corresponding to several days without the use of ensemble forecast systems (Marshall et al., 2014) because of the sensitivity to initial conditions (Frederiksen et al., 2004) and the model formulation (Pelly and Hoskins, 2003) too. Therefore, climate model evaluations that are run past the limit of deterministic predictability should not a priori expect a close consistency between weather-linked variations of global or regional temperature or precipitation between models and observations, except at the longer time-scales responding to external forcings (Gleckler et al., 2008; Santer et al., 2011). These differences reflect the well-known difference of the sensitivity of the results to errors in the initial conditions (predictability of the first kind) or to errors in the evolving boundary conditions (predictability of the second kind) (Chu, 1999; Lorenz, 2006).

There is not a universally accepted strategy for climate model evaluation, since it is well known that climate model evaluation and corresponding skill scores fundamentally depend on the target area, variable or intended application of the model evaluation study (Knutti, 2008). Some studies make a focus on the deterministic parts of the model simulations (basically, the seasonal cycle) (Boer and Lambert, 2001; Taylor, 2001). Other studies (oriented to the study of droughts or floods) tend to focus on extreme percentiles, since they are much more meaningful indicators of climate change (DeAngelis et al., 2013).

This study by DeAngelis et al (2013) is currently important for us, because it shows that models sometimes produce accurate values for the average of some climatic variable due to error compensation effects. They can, for instance, underestimate the frequency of high precipitation events and overestimate the frequency of low precipitation events. This points to the need to evaluate additional characteristics of climate model simulations beyond the mean value and standard deviation. Consequently, a few years ago, an index computed from the whole PDF of climatic variables was developed (Maxino et al., 2008; Perkins, 2007). It compares two PDFs and computes the minimum value of both PDFs at every abscissa. The area below this minimum represents the area below both PDFs. As such, for a perfect model, its value would be one, if both PDFs matched perfectly. This PDF-index or index of distributional agreement is the one that we will generalize to the multidimensional case in this contribution. The PDF-index analyses the

correspondence of the whole PDF both from a model and observations (Maxino et al., 2008; Perkins et al., 2007). The one-dimensional PDF-index has very often been used in the literature through the last years (Brands et al., 2012; Errasti et al., 2011, 2013; Fu et al., 2013; Maxino et al., 2008; Perkins et al., 2007; Schwalm et al., 2013; Ylhaïsi and Räisänen, 2013 to name a few). In this contribution we propose its extension to multiple dimensions, thus allowing to compare several features of climate or environmental models at a single step. The use of the PDF-index shows some advantages with respect to other approaches, in the sense that it samples the full PDF of the climatic variables. Therefore, the PDF-index is a very good index for the overall evaluation of the agreement between climate models and observed climate. However, there are other shortcomings, such as the fact that the analysis in terms of PDFs does not consider the time sequence of events, and the number of frost days or the number of continuous days without precipitation are important in terms of impacts (Brands et al., 2012). The PDF-index gives less weight to the tails of the distribution and, it is thus not adequate as the single index for the analysis of extremes (Brands et al., 2012).

In the case of the papers mentioned previously, the PDF-index is computed by means of unidimensional PDFs. However, in several cases, studies using the PDF-index or other scores (Dessai et al., 2005; Reichler and Kim, 2008) evaluate the skill of climate models according to several variables that may be of interest for the impact community. The most obvious instances might be precipitation and temperature, but if the scientists are interested in downscaling strategies, other variables such as geopotential height or sea level pressure appear very often (Brands et al., 2011; Errasti et al., 2011, 2013; Fu et al., 2013; Maxino et al., 2008; Radic and Clarke, 2011). In these previous references, the skill of the models is computed on a per-variable basis by means of univariate diagnostics. Their final skill score is computed by aggregating individual per-variable evaluations either by simple averaging or ranking of skill scores. However, there is currently a lack of universally accepted way of performing this combination of scores for different variables and the methodology that we propose in this contribution is aimed to fill this void, since a single index of distributional agreement is returned from the multidimensional PDF.

The main objective of this contribution is, therefore, to develop a methodology that can be applied to get a multidimensional score that allows to evaluate in a single step different variables from climate simulations against observations. In order to explain the advantages derived from using a multidimensional approach, we show a simple example derived from a synthetic dataset. We have created three synthetic datasets, G1, G2 and G3, derived from two-dimensional gaussian

distributions. For each case, the gaussians are centered $\mu_i = 0$ but the corresponding covariance matrices used to create them are given by $S_1 = S_2 = \begin{pmatrix} 1 & 0.75 \\ 0.75 & 1 \end{pmatrix}$ for G1 and G2, whilst for G3, the third gaussian, the covariance matrix is given by $S_3 = \begin{pmatrix} 1 & -0.75 \\ -0.75 & 1 \end{pmatrix}$. It can be seen that, despite the difference in the structure of the distributions of points (Figure 1, left), the univariate PDFs and corresponding indices show a good agreement (Figure 1, middle and right and Table 1), even though the distributions are different. This is quite an artificial example, but it illustrates the point that some parts of the PDFs close to the diagonals can be projected onto similar areas over the axis when using unidimensional indexes of agreement, masking the differences between the PDFs of the model and the observations. Therefore, it is interesting to analyse the full structure of the multidimensional PDF, since it yields a realistic difference between the score corresponding to G1 versus G2 (good agreement) and G1 versus G3 (bad agreement).

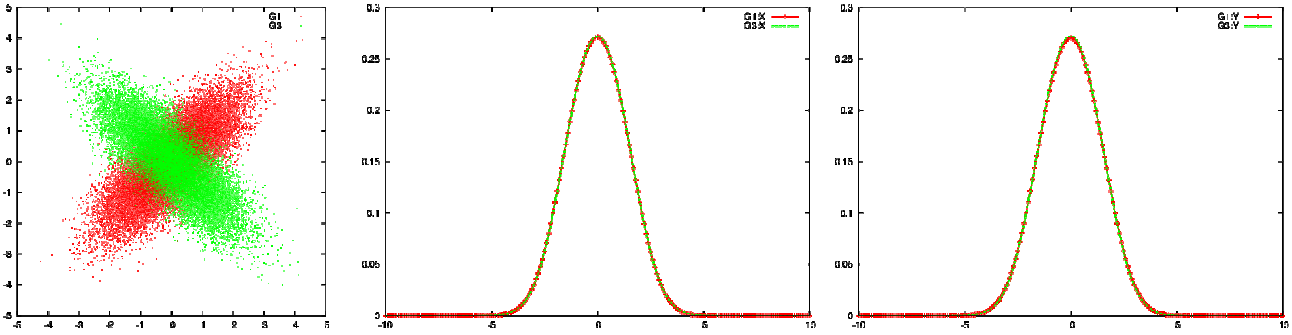

Figure 1. Points created from G1 (red) and G3 (green) distributions (left), univariate probability distributions corresponding to the X variables (middle) from G1 (red) and G3 (green) and univariate probability distributions corresponding to the Y variable from G1 (red) and G3 (green).

Table 1. Indices of distributional agreements for points derived from known gaussians using univariate scores for X and Y variables or two-dimensional scores.

		Univariate score		2D score
		G1-X	G1-Y	G1
G2	X	0.998		0.999
	Y		0.999	
G3	X	0.998		0.463
	Y		0.997	

To fill in the gap illustrated by the example, we generalize the PDF-index by Perkins et al. (2007) to *n*-dimensional phase spaces with the final aim of allowing an easy multi-criteria evaluation of models. That way, a single PDF-based index can group the performance of the models according to the multidimensional phase-space spanned by all the variables chosen for the evaluation of the model. In order to make it easier for other researchers to use this methodology, we present an implementation of this multidimensional extension by means of a set of tools that can properly address the computational problems that appear when making kernel-based estimations of PDFs with massive datasets. In the case studies that we show in this paper, we compute several realizations of the PDF-index using one to four dimensional phase spaces with up to 13 000 points for every particular model/realization. This is very intensive computationally, and for this reason we feel that an efficient implementation of the estimation of multidimensional PDFs could be of great help for researchers in this area. Thus, we have developed a general-purpose tool to compute kernel-based multidimensional PDF estimations that runs on state-of-the-art multi-core processors. Our proposal has two main characteristics: (1) a fine-tuned algorithm to calculate the PDF that minimizes the number of computations and (2) a parallel implementation of this algorithm that allows it to efficiently run in multi-core processors.

The method is applied to two different case-studies. The first case study corresponds to a realistic application of climate model evaluation (Errasti et al., 2013). Their results are re-analyzed using this new methodology. Additionally, a sensitivity of the results to the selection of the bandwidth parameter is carried out. Finally, the robustness of the results provided by the method to the existence of biases in the models is also studied. The second case study corresponds to the analysis of the performance of a coupled atmosphere-ocean reanalysis in reproducing the global scale Sea Surface Temperature and Sea Surface Height and it corresponds to a higher-dimensional problem. This second case study will be used to stress two of the merits attributed to the proposed methodology. On the one hand, this example in the context of the physical oceanography will demonstrate the wide range of the applicability of the methodology. On the other, the combination of variables with different physical dimensions will illustrate the ability of the method to process multivariate data and its ability to be applied to a large family of environmental models.

The remaining of this paper is structured as follows. Section 2 presents the materials and methods used in the paper. Results are shown in section 3. The discussion is presented in section 4, and the paper finishes with conclusions in section 5.

2 Material and methods

2.1 Data representing the daily seasonal cycle of zonally averaged temperature from global climate models and reanalysis.

For the first case study shown in this paper, we select a reduced dimensionality representation of the daily seasonal cycle of temperature that we already analyzed in Errasti et al. (2013). We analyze temperature of the air at the surface (TAS) daily data from seven models (20C3M simulations) of the CMIP3 experiment (Meehl et al., 2007) that were used by the Fourth Assessment Report of the IPCC (Randall et al., 2007). The models used are the BCCR-BCM2.0, GFDL-CM2.0, GFDL-CM2.1, MIROC3.2-HR, MIROC3.2-MR, MPI-ECHAM5 and MRI-CGCM2.3, and the basis for the selection of this subset of models and their characteristics can be found in Errasti et al. (2013). The same procedure is used for TAS data from ERA40 (Uppala et al., 2005) and NCEP/NCAR Reanalysis 1 (Kalnay et al., 2996), referred to as NCEP onwards. The TAS data from models and reanalyses were re-gridded to the same 2.5º x 2.5º grid by means of bilinear interpolation, since this was the coarsest grid used by any of the reanalysis used (the one used by NCEP). To reduce the dimensionality of such a gridded dataset, the TAS data were zonally averaged and projected onto Legendre polynomials that constitute an adequate basis over the sphere. Those have very often been used as a basis in one-dimensional energy balance models (North et al., 1981) and are the basis used for the meridional components of spherical harmonics used in spectral decompositions of the equations of motion (Washington and Parkinson, 2005). The Legendre polynomials are orthonormal in the set of functions over the continuous interval [-1, 1], but their corresponding discrete-grid counterparts are not orthogonal. For this reason, we applied the Gram-Schmidt orthogonalization procedure to obtain the leading discrete orthogonal $P_0(\mu)$, $P_1(\mu)$ and $P_2(\mu)$ Legendre polynomials. In the previous equations, $\mu = \sin(\theta)$ refers to the sine of latitude. The zonally averaged TAS profiles have been projected onto the orthogonal discrete $P_0(\mu)$, $P_1(\mu)$ and $P_2(\mu)$ Legendre polynomials and this has provided us with the corresponding time-varying coefficients $c_0(t)$, $c_1(t)$ and $c_2(t)$. Due to the meridional shape of the orthogonal discrete polynomials shown in Figure 1 in Errasti et al., (2013), the physical meaning of the temporal expansion coefficients can be easily understood. The coefficient $c_0(t)$ describes the seasonal evolution of global-mean temperature linked to the different distances from the earth to the Sun corresponding to the apogee or the perigee positions, $c_1(t)$ describes the seasonal evolution of summer-winter from one Hemisphere to the other and $c_2(t)$ describes the TAS differences between

the Equator and the poles. These $c_0(t)$, $c_1(t)$ and $c_2(t)$ coefficients represent the daily TAS seasonal cycle with a root-mean square error of the daily values ranging from 6.3 K (GFDL-CM2.0) to 8 K (MIROC3.2-HR) for the whole dataset. They represent 5%, 76% and 3.72% of the total variance of the zonally averaged TAS data for the case of ERA40 and similar values for the climate models and NCEP reanalysis, as extensively discussed by Errasti et al. (2013).

The use of two different reanalysis (ERA40 and NCEP) allows us to quantify the sensitivity of the results to the uncertainty of the temperature field. We consider that this uncertainty is given by the different values that we get from two reanalysis. Two different state-of-the-art evaluations of the TAS fields by two different reanalysis systems provide realistic representations of the surface temperature field, but the differences between them reflect the uncertainty in our knowledge of the detailed characteristics of the problem at hand. Both the 20C3M simulations and the reanalysis data cover the period 1961-1998 on a daily basis. The idea of using these data is to be able to compare the results of our multidimensional PDF index with already published results (Errasti et al., 2013) that used alternative techniques in terms of the skill score of the models. In Errasti et al. (2013) the PDF indices computed were one-dimensional and some diagnostics also involved the root mean square error. In this paper, we revisit the topic from a multidimensional point of view. For every *i*-th model, we have a three-dimensional vector of points $c_i(t) = (c_{0i}(t), c_{1i}(t), c_{2i}(t))$ that describes a trajectory in the phase space of the truncated daily zonally averaged air temperature at the surface. These data provide us with an interesting case study of the application of the new algorithm to a previous problem of evaluation of models where the multidimensional evaluation tool was not applied.

2.2 Multivariate oceanographic data.

Two datasets containing joint values for the Sea Surface Temperature (SST) and Sea Surface Height (SSH, relative to the geoid) are also analyzed as a second case study. These datasets are used in the second part of the Results section to illustrate the potential application of the proposed methodology in a multivariate application that not only combines variables with different physical dimensions (e.g. SST and SSH), but also different variables resulting from different truncations of the same physical field with the same physical dimensions (e.g. different projections of the SST field). Additionally, the interest of this second case study is that it also shows the potential application of the methodology in the case of other environmental applications not restricted to climatology.

Gridded global coverage data for the SST and SSH variables over the 1993-2012 period are used. The two datasets include: a reprocessed Level 4 (gridded, gap-free) product based on the merging by means of optimal interpolation of satellite and in-situ data called ARMOR-3D (Guinehut et al., 2004; Guinehut et al., 2012) and the coupled atmosphere-ocean CFSR reanalysis (Saha eta al., 2010). CFSR data covers the 1993-2010 part of the period considered here. For the 2011-2012 period CFSR is completed with data from the CFSv2 (Saha et al., 2014) analogue product. In the following the CFSR and CFSv2 product will be referred simply as CFSR.

To allow inter-comparison of the datasets they are averaged to weekly data and re-gridded to a common 0.5º x 0.5º resolution grid with an identical land-sea mask (following the minimal weekly time-frequency of the ARMOR-3D dataset and the minimal 0.5º x 0.5º spatial resolution of the CFSR dataset). Like in the case of the TAS, it is necessary to reduce the dimensionality of the SST and SSH variables in the datasets because a global dataset at a 0.5º x 0.5º horizontal resolution yields more than 2.0 x 10^5 grid-points, i.e. variables in the phase-space. This makes impossible to apply the proposed methodology to the raw data. In the case of the first case study (TAS) the dimensionality reduction is conducted by means of Legendre polynomials of zonally averaged values. In this case, the truncation is conducted by means of a Principal Component Analysis (PCA) (von Storch and Zwiers, 1999; Wilks, 2006) that extracts the Empirical Orthogonal Functions (EOFs) and the Principal Components (PCs) from the gridded anomalies (obtained after subtracting the time-mean). The gridded anomalies have been weighted to take into account the reduction of the grid-area with increasing latitude of a regular latitude-longitude grid. Therefore, the values at each grid point have been multiplied by the square root of the cosine of latitude (von Storch and Zwiers, 1999; Wilks, 2006). During this truncation process, orthonormal EOFs are used and, therefore, the units and the variances are retained by the new variables or PCs. For the sake of simplicity, the first two PCs will be retained for the SST (T1, T2) and SSH (H1, H2), making a total of 4 variables when all of them are combined. The percentage of total variances accounted for by those PCs are 83% (SST) and 19% (SSH) for the ARMOUR-3D dataset and 85% (SST) and 34% (SSH) in the case of CFSR.

2.3 Evaluation of the multidimensional PDFs, optimal bandwidth and evaluation of the multidimensional indices of PDF agreement.

The methodology suggested in this paper follows three steps that we describe in this order:

1. Compute a PDF by means of kernel estimates for the multidimensional case.
2. Identify the optimal bandwidth to be used by the estimation of the multidimensional PDF.

3. Find the amount of space common to the multidimensional PDFs of every model and observations.

From the point of view of the application, the first step would be to compute the optimal bandwidth, next, compute the multidimensional PDFs and, finally, compute the scores of the models' and observations' multidimensional PDFs. However, from the point of view of the description of the algorithm it is better described using the previous order because, in order to fully understand the way the optimal bandwidth is computed, the way the PDF is computed must be considered at the beginning.

A different program has been implemented for each step, which will be described next. Finally, note that the main contribution of this paper is not the use of these steps, but in the extension to multiple dimensions of the PDF-score. This extension required completely new programs, or important modifications to previously available codes. To the best of our knowledge, no other tool set is available to provide equivalent functionality with similar computational performance.

2.3.1 Multidimensional kernel density estimation

The first program (mpdfestimator) performs an estimation of the multidimensional PDF in a multidimensional space. It takes as input a file containing all the observed points and optionally, a bandwidth value and the parameters defining the evaluation space (minimum, maximum and increments for every dimension). If the bandwidth value is not provided, the default corresponding to a multidimensional gaussian distribution with the same sample size is applied (Silverman, 1986).

Program *mpdfestimator* computes the PDF as $\hat{f}(x,h) = \frac{1}{nh^d}\sum K\left(\frac{x - x_i}{h}\right)$ where n is the number of observations, K is the kernel function, d is the number of dimensions of the dataset, h the bandwidth value, x_i is a vector containing each observed point, and x is a point of the space where the PDF is being evaluated (evaluation point onwards). The PDF is estimated in a user-defined space called *evaluation space*.

According to Silverman (1986), asymptotically, there are no differences between the different kernels at hand (Gaussian, triangular, Epanechnikov, etc.). Moreover, he states that it is desirable to base the choice of the kernel on other considerations, such as the degree of differentiability required or the computational effort involved. Other references also support the fact that the sensitivity of the results to the kernel chosen is small (Ahamada and Flachaire, 2010) and that the Epanechnikov kernel is very efficient (Scott, 1992). In our program, since the Epanechnikov kernel is bounded, we will take advantage of this boundedness to design a computationally efficient proposal.

The Epanechnikov kernel function is defined as: $K(\Delta x) = \frac{(d+2)}{2c_d}(1 - \Delta x^T \cdot \Delta x)$ where c_d represents the volume of the d-dimensional sphere of unit radius and $\Delta x = x - x_i$ is the vectorial difference between an observed x_i and an evaluation x point. The use of a simple euclidean norm would be misleading in case the variables used to define the phase space were characterized by very different variances. In order to avoid this, the algorithm computes the spherically symmetric principal components y_i that can be derived from the x_i observed points. Note that the data are also initially centered to be able to compute principal components. There is no filtering of the main principal components in this step. If a reduction of dimensionality is required by the user, it must be performed by the user before calling this program. We, therefore, assume that the covariance matrix of the input data is of full rank. The fact that we use spherically symmetric principal components means that variables expressing more variance of the input data are given more importance in the computation of the kernel and, as such, the same bandwidth (*h*) can be used for all the dimensions. This is equivalent to using a Fukunaga-like estimator (Silverman, 1986). The same linear mapping that is used onto the observed x_i points to compute the corresponding spherically symmetric principal components y_i is applied to the centered evaluation points x yielding a set of scaled evaluation points y. The difference vector and the kernel function are computed by using $\Delta y = y - y_i$. Therefore, the PDF is actually defined in a set of coordinates y and, after the PDF has been evaluated, it is transformed back to the original evaluation grid x. This means that it is divided by the Jacobian of the transformation (square root of the determinant of the covariance matrix) from the original evaluation space x to the new evaluation space y in order to recover the proper normalization of the PDF (Menke and Menke, 2012). If the data have been centered when the PDF was being computed, the axes defining the PDF are again translated at this point according to the original mean that has been computed before storing the results in the output netCDF files. At this point, it is convenient to stress that the resulting netCDF files are stored around the original average of the input dataset using the physical variables corresponding to the original phase space. As will be explained below, the internal use of principal components will also be a critical part in the development of the bootstrap for the multidimensional case. Additionally, it has to be mentioned that in multivariate cases, such as the comparison between SST (K) and SSH (m) that we present in the second case study, the metric that we use to evaluate the kernel using radially symmetric principal components allows us to use a non dimensional *h*, thus properly combining multivariate datasets. Thus, on the following pages, when *h* is mentioned, it refers to a non dimensional

parameter.

A common approach to compute the estimate of the PDF, valid for any kind of kernel, is to make a loop that traverses all the evaluation points of the space, and calculates the kernel function for each <observation point, evaluation point> pair. This approach, evaluation point-based (EPB onwards), is shown in Listing 1.a. Its complexity is $O(k_d \cdot m \cdot n)$, being k_d a constant related to the dimensionality of the dataset, *m* the number of evaluation points and *n* the number of observed points.

```
a) Evaluation point-based
      estimation (EPB)
for each EvaluationPoint x {
      den_x = 0

      for each ObservedPoint x_i {
            den_x += density(x,x_i)
      }
}
```

```
b) Observed point-based
      estimation (OPB)
for each EvaluationPoint x {
      den_x = 0
}
compute Observation Influence Area A
for each ObservedPoint x_i {
      for each EvaluationPoint x in A {
            den_x += density(x,x_i)
      }
}
```

Listing 1. Approaches to computing *PDF estimation expressed as pseudo-code.*

In some cases, the kernel that defines the density of each sample is bounded and it affects only a small subset of the evaluation points in the evaluation space. In these cases, using the EPB approach is inefficient, as most of the evaluation points lie outside the area of influence of the kernel (but a calculation is required), being thus the corresponding density zero. Thus, as we have chosen a bounded kernel (the Epanechnikov kernel), we have defined a way to estimate the PDF that reduces the computational complexity of the previous approach, aiming to minimize the execution time of *mpdfestimator*.

Our approach, observation point-based (OPB onwards), performs a loop traversing the observation points, computing for each, the density over the evaluation points affected by the influence area of the kernel. This method requires to identify the set of evaluation points inside that influence area, which can be done by means of geometrical equations. As a visual example, we depict in Figure 2 the influence area of a kernel as a grey ellipse. In our program, we calculate a square shaped *bounding box* around each observation point, which includes some evaluation points outside the area of influence, but it is easier to process by the program.

A simplified OPB is described in Listing 1.b and its computational complexity is $O(k_d \cdot m \cdot a)$, being k_d a constant related to the dimensionality of the dataset, *m* the number of observed points, and *a* the number of evaluation points inside a bounding box (generally much smaller than *n*).

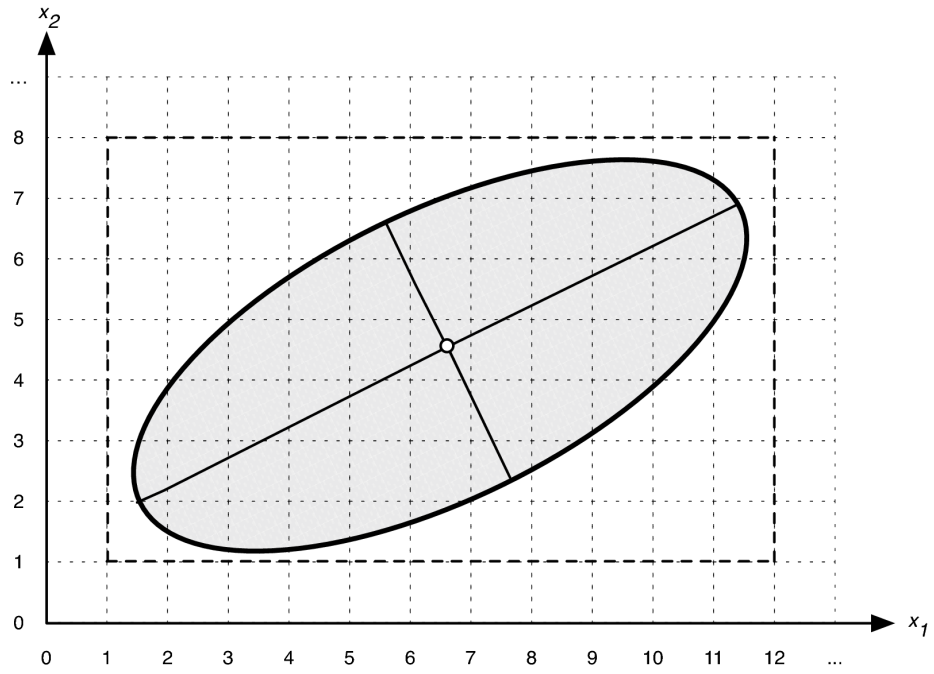

Figure 2. Example of the influence area around an observed point in a 2D space, identifying all the grid points affected. The ellipse represents the actual influence area in the physical space, while the dashed rectangle shows the prismatic bounding box around it.

mpdfestimator (as well as the remainder programs) has been implemented using the C programming language. All the linear algebra routines used are implemented through calls to the publicly available MESCHACH library (Stewart and Leyk, 1994). The PDF is written to a netCDF file in the physical evaluation space x requested by the user with the PDF centered and scaled according to the original data. This means that the user is free to evaluate PDFs from models and observations that are (are not) centered at the same *n*-dimensional average points (biased or unbiased datasets). We recommend, as shown below, that datasets are always bias-corrected before performing this analysis, but the system does not enforce the user to do so, althought it warns the user about this condition.

In the first case study used in this paper the phase-space is tridimensional, but the implemented program allows the user to go up to any dimension (starting from one), as shown by the second case study used in this paper. The one-dimensional case is also covered by our program even though there are other implementations that can cover the one dimensional case. The novelty of our contribution lays on the generalization of the one-dimensional score to higher dimensions.

In addition, standard OpenMP programming directives (Dagum and Menon, 1998) have been also included with the aim of exploiting the multi-core capability of present computers. The set of observed points will be equally distributed amongst the processors for their computation in parallel. This way, the workload is split among the available processors, reducing the execution time (almost) linearly to the number of cores used.

2.3.2 Selection of the optimal bandwidth

The second step that we describe (although it should be the first step in the application of the programs) is to find the optimal bandwidth value for the PDF computed in the first step. It is well known that the computation of the optimal bandwidth to be used in PDF estimations using kernels is a critical step in obtaining reliable PDFs. There are two major strategies for the determination of the optimal bandwidth (Scott, 1992; Slverman, 1986): cross-validation (Duong and Hazelton, 2005) and smoothed bootstrap (Faraway and Jhun, 1990).

The cross-validation approach leads to the convolution of the kernel with itself, a very tough mathematical problem for the Epanechnikov kernel with an open number of dimensions. It is usually solved by means of gaussian multiplicative kernels (Duong and Hazelton, 2005), but this wouldn't allow us to use the OPB approach explained in section 2.3.1 above. In our case, we have selected the use of smoothed bootstrap estimates of the optimal bandwidth, since it simplifies the generalization of the solution to an open number of dimensions in the multidimensional case for the non-multiplicative Epanechnikov kernel we are using. In order to produce the new estimations in the multidimensional case we take advantage of the fact that the kernel is spherically symmetric in the space corresponding to the spherically symmetric principal components. Thus, the same strategy used by univariate kernel estimations is used for every direction in the space spanned by the spherically symmetric principal components (Silverman, 1986). Surrogate samples are created in this space, and this procedure guarantees that the structure of the covariance matrix is properly preserved.

Following Faraway and Jhun (1990), we use a smoothed bootstrap procedure to estimate the squared error between two estimates of the PDF. The smoothed estimate starts from a reference evaluation of the PDF $\hat{f}(x,h_0)$ computed using a reference bandwidth h_0. Then, several estimations $\hat{f}_n(x,h)$ of the PDF are performed at varying values of the bandwidth parameter *h* and a number of $n=1\dots N$ realizations for every *h.* The bootstrap program checks the error between the "reference" PDF used in the smoothed bootstrap procedure and the actual bootstrap samples by evaluating the squared error $\varepsilon_n(h)=\int \left(\hat{f}(x,h_0)-\hat{f}_n(x,h)\right)^2 dx$. Then, the bootstrap-derived distribution of the squared errors is used to infer minimum, maximum, median, $P_{0.025}$ (2.5%) and $P_{0.975}$ (97.5%) percentiles of squared error for every value of *h*. This information is reported to the user at every *h* value. The *h* value producing the lowest values of the error estimates (we use the median of $\varepsilon_n(h)$ in the case studies in this paper) is the one selected as the optimum bandwidth.

This procedure has been implemented in the *mpdfestimator_bootstrap* program. It takes as input all the observed points and optionally (1) a reference bandwidth value, (2) a range of bandwidth values to be evaluated, (3) the boundaries of the evaluation space, and (4) the number of repetitions for the random sampling. If (1) is missing, the default corresponding to a multidimensional gaussian distribution with the same sample size is applied. If (2) is missing a range (+/- 20% around (1), with a step such that the maximum bandwidth interval is divided in 10 subintervals) is defined. In case (3) is missing, *mpdfestimator_bootstrap* defines a range that ensures a space that surrounds all the observed points. Finally, if (4) is missing, 500 realizations are performed. The program generates as output squared errors for each of the provided bandwidth values. The pseudo-code is shown in Listing 2.

```
Compute the PDF for the reference bandwidth h0
for each h in the range [hmin,hmax]{
      for iter=1 to max_repetitions{
            generate a random sub-sample S
            compute PDF for S
            compute squared error
      }
      generate statistics of squared error
}
return statistics
```

Listing 2. Pseudo-code for the *mpdfestimator_bootstrap* program.

2.3.3 Computation of the PDF score

The final step of the methodology is to compute the PDF score. Once the user has computed the PDFs (by means of the *mpdfestimator* program) corresponding both to the model and the observations using the optimal bandwidth value reported by *mpdfestimator_bootstrap,* program *mpdf_score* has to be executed to get the PDF score against the reference model.

The program *mpdf_score* takes as input two *n*-dimensional PDFs stored as netCDF files, generated for the same domain by the first program *mpdfestimator*, and provides as output a PDF-index *S* by means of the following equation (adapted in this case for a three-dimensional example): $S=\sum min\left(Z_{ijk}^{o},Z_{ijk}^{m}\right)dx_i\,dx_j\,dx_k$, where Z_{ijk}^{o} and Z_{ijk}^{m} refer to the evaluation of the PDF from observations and the model, respectively. Please note that for higher dimensions, the extension is straightforward.

The equation closely follows the one used by Perkins et al. (2007) or Maxino et al. (2008), but has been extended in this case for its use with a PDF defined in a multi-dimensional (*n*-dimensional) space. Additionally, when working in several dimensions, the volume of the *n*-dimensional interval where the PDF is being computed must be taken into account for normalization purposes, and so, the dx_i , dx_j and dx_k terms account for the fact that the range of the different variables can be very different (the average standard deviations of the coefficients in our first case study are 1.9 K, 9.7 K and 2.1 K). The program warns the user in case the bias for any of the dimensions is greater than 5% of the standard deviation of that variate.

2.4 Representation of marginalized PDFs for the interpretation of results.

Finally, even though it is not part of the methodology we propose, in order to be able to identify the differences in the index corresponding to individual models and for illustration purposes of the results corresponding to the first case study, we have computed marginalized $\hat{f}_{2D}(c_i,c_j,h)=\int \hat{f}(x,h)dc_k, i\neq j\neq k$ two-dimensional PDFs and projected them onto the *i-j* C0-C1, C0-C2 and C1-C2 planes, after marginalizing *k* axes C2, C1 and C0, respectively. This will allow us to show that using a single multidimensional score is better than using a set of unidimensional scores. We only present marginalized PDFs for the first case study in the paper and, for the second case study we just collect the aggregated values of the score in a table.

3 Results

3.1 Application to climate model simulation of the daily cycle of surface temperature

Figure 3 shows the evolution of the median of the squared errors and the 95% confidence interval computed from the bootstrap analysis corresponding to the ERA40 data when the reference PDF is computed with two conservative estimates of bandwidth (h_0=0.8 and h_0'=0.67) against the bandwidth that would correspond to the same sample size for a gaussian PDF, 0.637. It can be seen that the bootstrap estimate suggests a slightly lower value (0.55-0.60) of the bandwidth parameter than the one that would correspond to a gaussian multidimensional PDF. As will be identified in the marginalized PDFs later, this is to be expected, since the zonally averaged surface temperature is very non-normal and periodic, so that several fine scale features of the PDF must be resolved, and they can only be properly resolved if the bandwidth is not very high. Therefore, in the following steps an optimum bandwidth of h=0.6 will be used unless otherwise explicitly stated.

In order to test the sensitivity of the classification to different values of the bandwidth used, Table 2 presents the results of the multidimensional PDF scores for different values of the bandwidth parameter (every model is centered and checked against ERA40). For the optimum bandwidth (h=0.6), the best model available is the alternative reanalysis that is used in this study (the NCEP). This is something that we expected from the beginning, since both reanalyses are based on observations. This result supports the use of the method, since the method yields better results for alternative observation-based reanalyses. MIROC3.2-MR and HADGEM1 are the models that follow. Some of the model runs differ only on the initial conditions and most of them are grouped together, with the exception of HADGEM1. The interpretation of this result is that the variability of the index to the use of different initial conditions is very low, as should be expected. MIROC3.2-HR, GFDL, ECHAM5 and BCM2 follow the previous models. The ranking finishes (for the subset of models and diagnostic variable used in this study) by the five random runs corresponding to the MRI model. All the runs corresponding to MRI are grouped, with low values of the score that do not mix with values corresponding to the rest of the models. It seems, therefore, that the intra-ensemble variance is in general (without the exception of MIROC3.2-MR and HADGEM1) smaller than the inter-model variance of the score. In general, the main characteristics of these results are robust even with changes in the bandwidth that span a -33% to a +33% interval from the optimum value found by means of bootstrap. The models that show the highest (lowest) performances with the optimum value of the bandwidth continue showing a similar performance for higher or lower values of the bandwidth. There are occasional excursions of a model to at most one alternative position up/down of the ranking, but, on the whole, models tend to maintain their relative ranks

even when the bandwidth is changed by a +/-33% relative change around the optimum value.

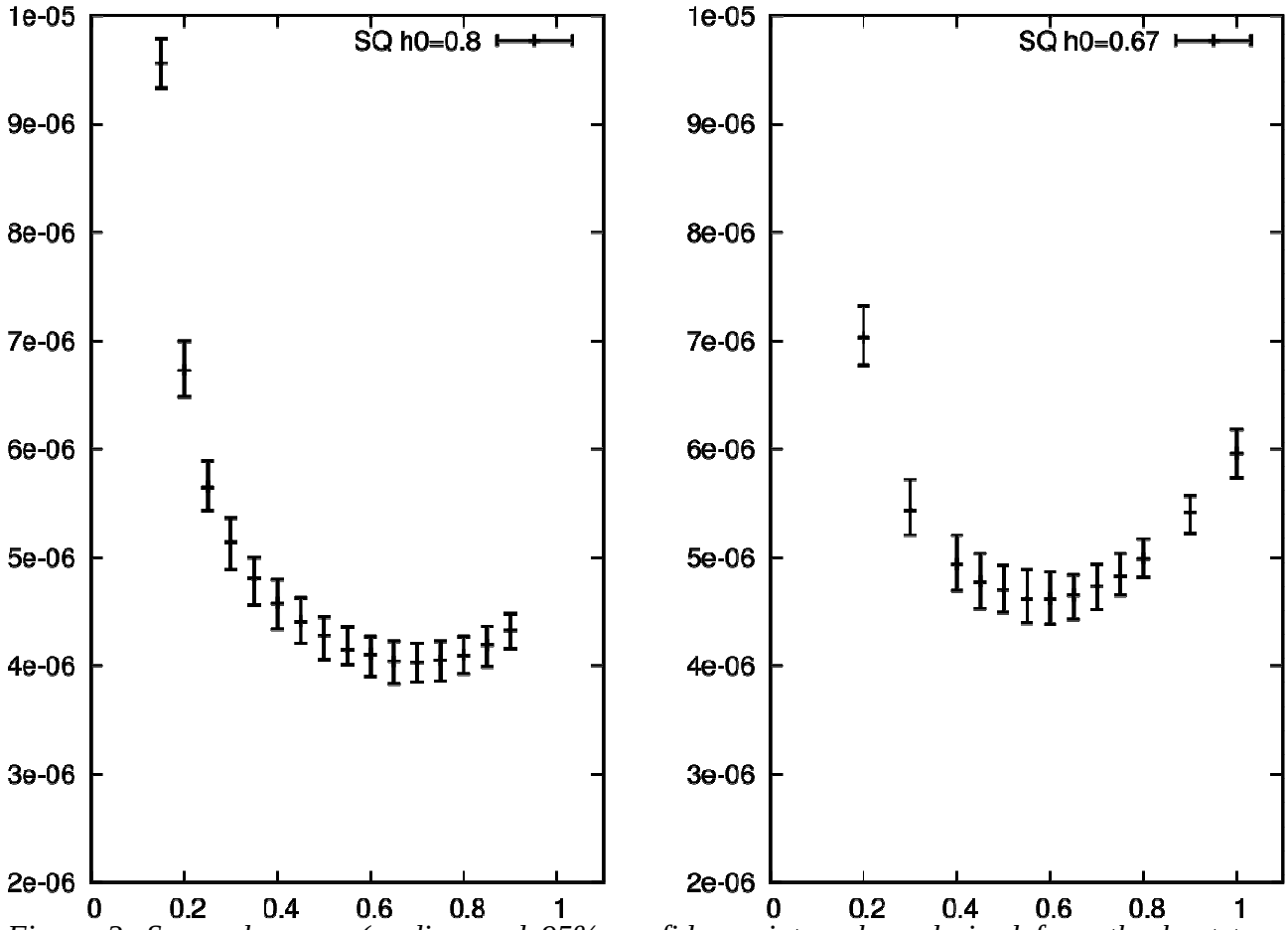

Figure 3. Squared errors (median and 95% confidence interval as derived from the bootstrap estimates) between the randomly generated PDFs and the reference PDF (left, h_0=0.8 and right, h_0=0.67) used for the generation of the smoothed bootstrap.

Figures 4, 5 and 6 show the plots of the marginalized PDFs for the case of the NCEP (contours) versus ERA40 (shaded), a model showing a high value of the score, MIROC-3.2-MR (contours) versus ERA40 (shaded) and a model with a lower score, such as MRI (contours) versus ERA40 (shaded). In order to show simple numbers in the plots and scales, values of the marginalized PDFs are multiplied by one thousand before plotting. Before computing the PDFs, the biases between every model and ERA40 have been removed by centering all the series.

Table 2. Values of the multidimensional ***S*** score corresponding to different values of the bandwidth parameter and associated rankings that would correspond to the models, when compared with ERA40 reanalysis data.

	h=0.4		h=0.5		h=0.6		h=0.7		h=0.8	
Model	S score	rank	S score	rank	S score	rank	S score	rank	S score	rank
BCM2.0	0.45	10	0.48	9	0.51	9	0.53	9	0.56	9
ECHAM5	0.45	9	0.47	10	0.48	10	0.50	10	0.51	10
GFDL-CM2.0	0.55	8	0.58	8	0.60	8	0.61	8	0.63	8
GFDL-CM2.1	0.58	7	0.60	7	0.62	7	0.64	7	0.65	7
HADGEM1	0.66	5	0.69	4	0.71	4	0.72	4	0.73	4
MIROC3.2-HR	0.62	6	0.65	6	0.67	6	0.70	6	0.71	6
MIROC3.2-MR-RUN01	0.66	4	0.68	5	0.70	5	0.72	5	0.73	5
MIROC3.2-MR-RUN02	0.69	2	0.72	2	0.74	2	0.75	2	0.77	2
MIROC3.2-MR-RUN03	0.69	3	0.71	3	0.73	3	0.74	3	0.75	3
MRI-RUN01	0.25	13	0.27	13	0.29	13	0.31	13	0.33	14
MRI-RUN02	0.25	14	0.27	14	0.29	14	0.31	14	0.33	13
MRI-RUN03	0.25	11	0.28	11	0.30	11	0.32	11	0.34	11
MRI-RUN04	0.25	12	0.27	12	0.29	12	0.32	12	0.34	12
MRI-RUN05	0.23	15	0.25	15	0.28	15	0.30	15	0.32	15
NCEP	0.80	1	0.81	1	0.82	1	0.83	1	0.84	1

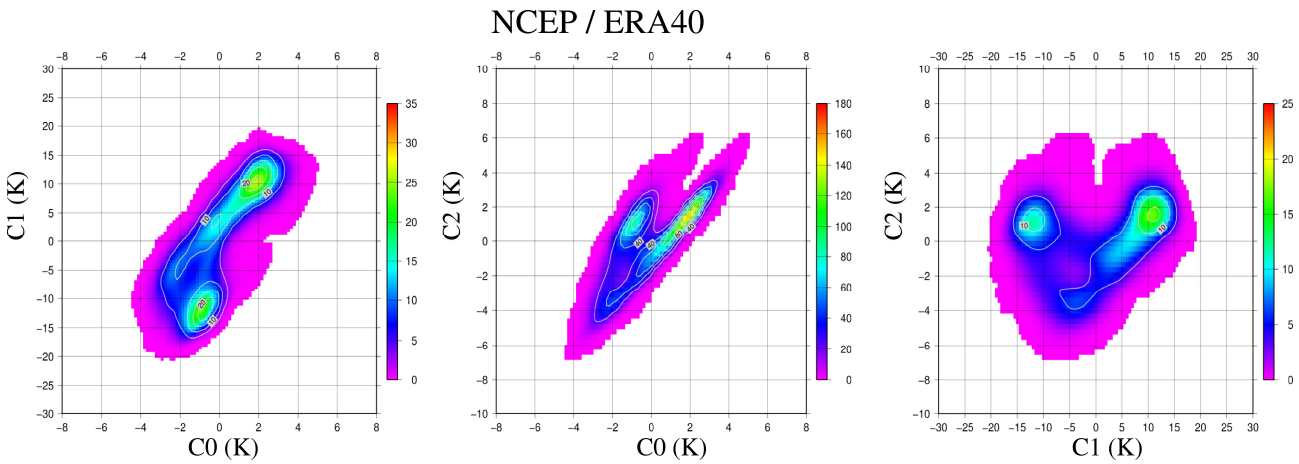

Figure 4. Marginal PDFs of NCEP (contour) and ERA40 (shaded) projected onto the planes defined by the C0-C1 coefficients (left), C0-C2 coefficients (middle) and C1-C2 coefficients (right). Values of the PDF have been multiplied by 1000 in order to improve the representation of numbers.

Figure 4, left, shows that on the C0-C1 plane, the PDF is clearly bimodal, as should be expected from a periodic deterministic signal such as the seasonal cycle of temperature. C0 represents the global average of surface temperature and C1 represents the difference in temperature between the Northern and Southern Hemispheres. The main clusters of the C0-C1 PDF appear aggregated around each Hemisphere's summer. NCEP values show a slightly warmer global temperature (C0) during Southern Hemisphere summer than the values shown by ERA40. Figure 4 (middle) shows that the amplitudes and phases of the mean global temperature (C0) and the equatorial bulge (C2) are similar in both reanalyses. On the C1-C2 plane, the marginalized PDF shows that the main difference between both reanalyses appears as a slightly higher difference of temperature between hemispheres (C1) in NCEP when the coefficient representing the equatorial bell (C2) is positive

(summer in the Northern Hemisphere). However, the PDFs generated by both reanalysis are extremely similar, as reflected in the high value of the S index between NCEP and ERA40 (0.82). This is something that we expected from the beginning, since they correspond to observational datasets.

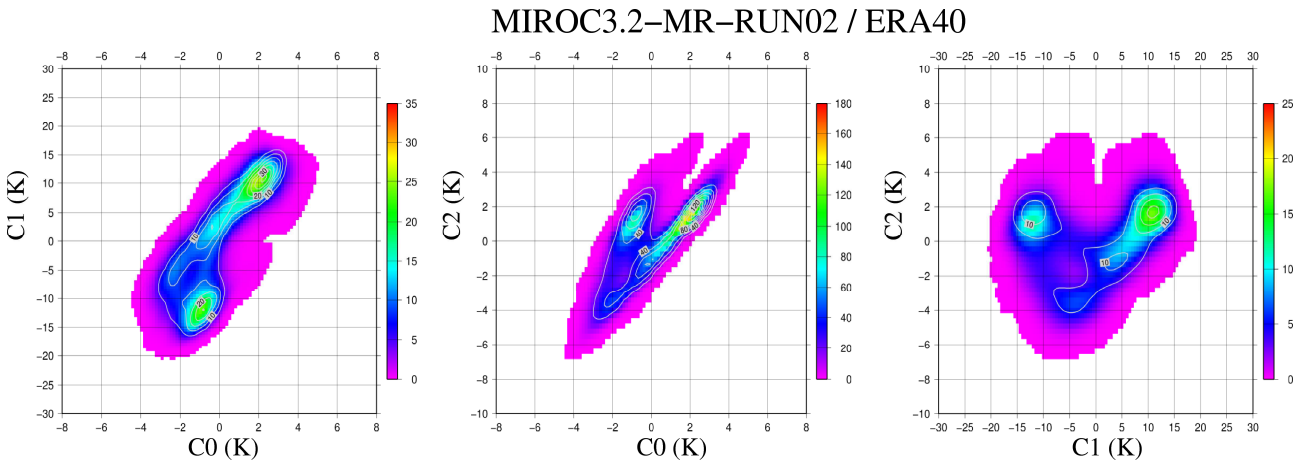

Figure 5. Marginal PDFs of MIROC3.2-MR (run 2, contours) and ERA40 (shaded) projected onto the planes defined by the C0-C1 coefficients (left), C0-C2 coefficients (middle) and C1-C2 coefficients (right). Values of the PDF have been multiplied by 1000 in order to improve the representation of numbers.

Figure 5 corresponds to the marginal PDFs for MIROC3.2-MR model (second run), one of the best CMIP3 models according to the metric selected in this study. Over the C0-C1 plane (left), there is quite a good agreement between both PDFs, since both clearly represent the bimodal structure of the PDF. However, the differences between MIROC3.2-MR and ERA40 are higher than in the previous case, both in terms of the location of the Northern Hemisphere summer and also in transitions between seasons that appear in the areas between the maxima in the marginal PDF. In the case of the C0-C2 plane (middle), the highest disagreement appears at the precise location of the maxima of the marginal PDFs, particularly during Northern Hemisphere summer. A similar diagnostic can be derived from the marginal PDF over the C1-C2 plane. Despite both marginal PDFs are clearly bimodal, slight differences exist at the placing of the PDF maxima. The equatorial bell (C2) in MIROC3.2-MR is stronger than the one in ERA40 during negative phases (Northern Hemisphere winter) of inter-hemispheric temperature differences (C1).

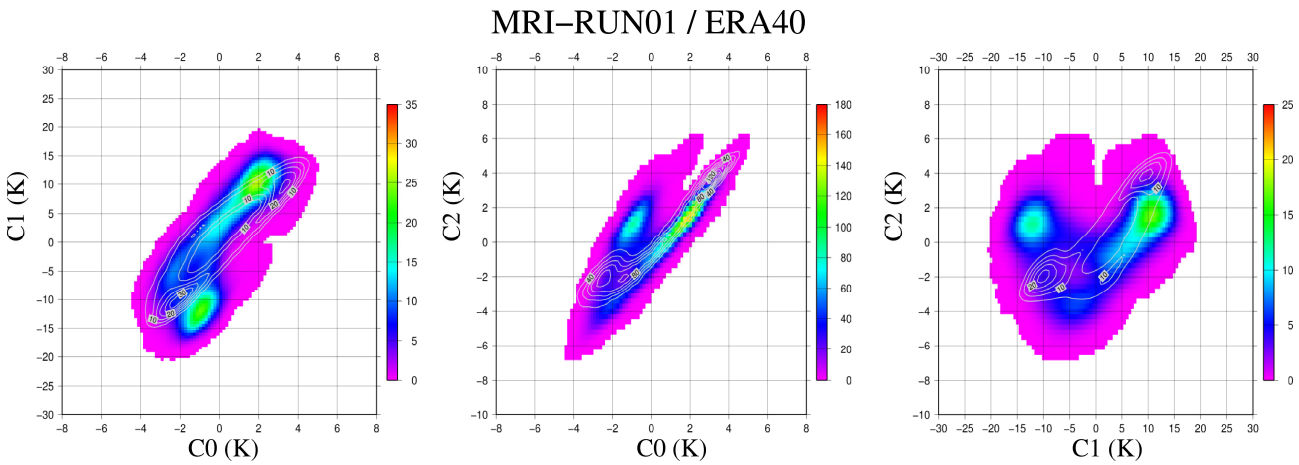

Figure 6. Marginal PDFs of MRI (run 1, contours) and ERA40 (shaded) projected onto the planes defined by the C0-C1 coefficients (left), C0-C2 coefficients (middle) and C1-C2 coefficients (right). Values of the PDF have been multiplied by 1000 in order to improve the representation of numbers.

Figure 6 corresponds to MRI (run 1) model. The evolution of the daily seasonal cycle of temperature in terms of C0 (global T) and C1 (inter-hemispheric temperature contrast, left) does not present a bimodal structure with the PDF maxima placed at the same points shown by the reanalysis. The transitions between summer and winter regimes happen through routes that do not correspond to the ones in the ERA40 Reanalysis. The structure of the marginal PDF for the C0-C2 plane is markedly different between MRI and ERA40, with the cold maximum in the PDF during summer in the Southern Hemisphere quite misplaced in the case of MRI. This is also apparent in the marginal PDF corresponding to the C1-C2 plane, where maxima of the PDFs do not appear neither on the same places nor even with the same phases.

Finally, Figure 7 shows that the index is very sensitive to the existence of a bias between the models and reference observations. In this case, the PDFs are computed without previously removing the bias between the surrogate model (NCEP data) and the observations (ERA40) and the S score index that we get between ERA40 and NCEP reanalyses is extremely low (S=0.075). The marginal PDFs show that in general there is a very good agreement in the structure of the 3D PDFs, but the center of masses of both PDFs are not located at the same places. The biases for every coefficient are not very high, considering their variances. The bias of the C0 component is 0.7 K (0.2% relative error), the bias in C1 is 0.4 K (7.5% relative error) and the bias in C2 is -0.34 K (-1.33% relative error). However, even such low values of the bias lead to a score index that could be interpreted as poor performance of the surrogate model (NCEP reanalysis) versus ERA40 due to the complex structure of the 3D PDF. However, this is a false impression that can not be defended if the spatial patterns of

the marginal PDFs are analyzed in detail. This means that the index should not be applied to model results that are biased against the reference observations. The existence of biases in the models leads to greater observational uncertainty when the model and observational datasets are not centered. The code does not force the centering of the datasets and, therefore, the user must take care of this when the dimensionality reduction stage of the data analysis is done. In particular, it is interesting to stress that, internally, when computing the *n*-dimensional PDFs, all the datasets are centered (each one using its *n*-dimensional average) before computing the corresponding spherically symmetric principal components. When the output netCDF files holding the PDFs are saved, the original units in the phase space of each dataset (model or observations) are recovered and the average is added to the anomalies derived from the PDF in the principal component space stored in the memory of the computer. Therefore, the key point here is that if there exists a constant bias between the model and the reference observations (first and second netCDF files passed to program mpdf_score), it could lead to very low values of the score despite the model representing properly the variability (anomalies). This means that the evaluation of the models in terms of a constant bias and the *n*-dimensional PDFs should be carried out as different steps.

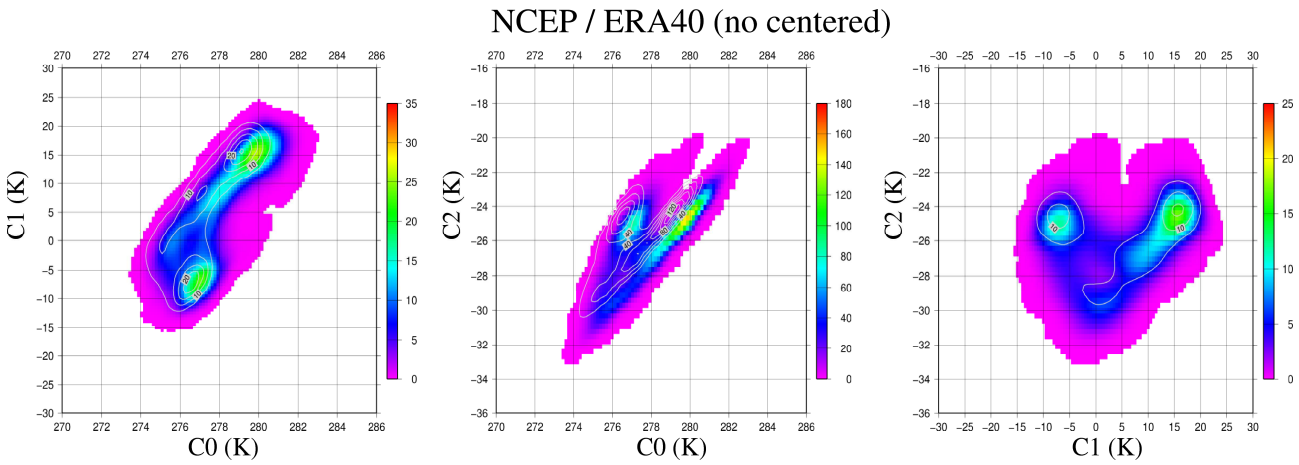

Figure 7. Marginal PDFs of non-centered NCEP (contours) and ERA40 (shaded) projected onto the planes defined by the C0-C1 coefficients (left), C0-C2 coefficients (middle) and C1-C2 coefficients (right). Values of the PDF have been multiplied by 1000 in order to improve the representation of numbers. The bias between both reanalysis has been retained.

From the point of view of performance, we have measured the execution time needed by each version of the program to complete the bootstrap procedure. On average, the serial OPB approach is 140 times faster than the serial GPB approach and, moreover, the parallel OPB program scales

linearly with the number of cores, being 4.3 times faster than its serial counterpart when using 4 cores. This means that an evaluation of a single model that takes approximately 22 days with the serial GPB program, can be executed in less than one hour using the most efficient and parallel implementation presented in this contribution. These experiments have been conducted in a desktop computer with an Intel i7 3820 processor (four cores, 3.6GHz, Hyperthreading enabled) with 8GB of RAM. Therefore, the use of this technique is not limited to the availability of specialized clusters or hardware that would limit its practical use.

3.2 Evaluation of Sea Surface Temperature and Sea Surface Height

The first two PCs of the global coverage weekly time-scale SST (T1, T2) and SSH (H1, H2) variables belonging to the ARMOR-3D (Guinehut et al., 2004; Guinehut et al., 2012) and CFSR (Saha et al., 2010; Saha et al., 2014) datasets will be used in the following to evaluate the second with respect to the former. This means that the ARMOUR-3D product (blended satellite and in-situ observation product) is the reference to evaluate the CFSR product (coupled atmosphere-ocean modelling product). Considering the first two PCs of each variable in the evaluation (T1, T2, H1, H2), the global-scale main variability modes of each variable are being take into account at a glance. As the seasonal cycle was not explicitly removed from the anomalies used to deduce the PCs, the four considered variables are almost completely related to the global-scale seasonal cycle (H2 contains some longer time-scale variability). Thus comparing combinations of different variables from CFSR with those of ARMOR-3D the capacity of the modelling product to jointly characterize different main global-scale variability modes (their seasonal cycles) is evaluated. For example, if T1 and H1 are considered at the same time (case T1H1) the capacity of CFSR to simulate the main global-scale components of the seasonal cycle of the SST and SSH variables is being evaluated in a single and multivariate score.

Table 3 shows the optimal h and the score obtained with the 6 analyzed cases going from the univariate T1 and H1 cases, the multi-dimensional univariate T1T2 and H1H2 (reserving the term multivariate to the cases with variables with different physical dimensions, i.e. Kelvins and meters) and the multivariate T1H1 and T1T2H1H2 cases. All variables have zero mean so no bias related issues will be observed in this case. Like in the previous case study on the TAS, the same three-step methodology was applied in this case: for a given row in Table 3, the optimal h using the bootstrap procedure is initially estimated. Next, the PDF using the optimal h is computed and, finally, the score (one dimensional, multidimensional or multivariate) is computed from the PDFs obtained from the CFSR and the ARMOR-3D variables.

The very high scores of T1 (0.98) and T1T2 (0.97) cases indicate that the model is reproducing well the global-scale SST seasonal cycle. The same stands for the SSH, as slightly lower but still high values are observed in the analogue H1 (0.84) and H1H2 (0.79) cases. This difference between the SST and SSH is not strange as it is only showing the fact that the modelling of the SSH involves many more processes (dynamic anomalies, steric anomalies,...) compared to that of the SST (mixed layer heat-budget and currents). The multivariate cases T1H1 (0.67) and T1T2H1H2 (0.42) indicate a loss of performance with growing number of the dimensions of the multivariate PDF, or in other words, with increasing complexity. Any error in any of the tested dimensions leads to a smaller value of the score. It can be seen that, although univariate scores point to good modelling performances, the multivariate approaches combining the same variables yield worse results. In this application a single model is being evaluated, but results show that differences appear when considering multivariate scores that discriminate between errors that appear through all the phase space.

Table 3: Optimal *h* and resulting scores obtained from the evaluation of CFSR (model) with respect to ARMOR-3D (observations) based on the first PCs of the SST (T1, T2) and SSH (H1, H2) from each of the products. Univariate unidimensional (T1 and T2), univariate multidimensional (T1T2 and H1H2), multivariate unidimensional (T1H1) and multivariate multidimensional (T1T2H1H2) cases are shown to illustrate the number of potential combinations that can be considered in the framework of the proposed methodology.

CASE NAME	h (optimal)	SCORE
T1	0.36	0.98
H1	0.54	0.84
T1T2	0.52	0.97
H1H2	0.69	0.79
T1H1	0.58	0.67
T1T2H1H2	1.04	0.42

4 Discussion

4.1 Comparison with the results of the evaluation of models using previous techniques.

The results presented in this contribution extend the previous analysis by Errasti et al. (2013) with a new methodology. Therefore, the obvious first part of the discussion is a comparison of the rankings

obtained by this methodology and the final rank obtained in the previous study. It can be seen that, for the most part, the aggregated rankings behave similarly in the current and previous studies (see Table 4). In general, the best models (NCEP, MIROC3.2-MR and MIROC3.2-MR) keep a good ranking also under the new metric proposed here. For the worst case (MRI-CGCM2.3), the same result holds, too. However, some changes appear in the relative rankings of models in the middle part of the classification. It must be kept in mind that the rankings in Errasti et al. (2013) also considered the root mean square errors between the seasonal cycles, so that the rankings can not be completely the same. Anyway, as a result of the new technique, we find a similar classification of models albeit with a single objective index without the need of a posterior averaging of individual scores or ranks. Therefore, we think that this methodology makes it easier and more objective to classify the models according to their performance when simulating several variables.

Table 4. Global rankings for the models according to their representation of the daily zonally averaged temperature in Errasti et al. (2013) and according to the methodology proposed in this contribution (3D-PDF).

	Errasti et al. 2013 rank	3D-PDF rank
BCM2.0	5	6
GFDL-CM2.0	7	5
GFDL-CM2.1	6	4
MIROC3.2-HR	2	3
MIROC3.2-MR	3	2
ECHAM5	4	7
MRI-CGCM2.3	8	8
NCEP	1	1

The technique presented in this contribution is slightly different of those carried out in other model inter-comparison studies found in the climate literature. In some studies estimates of simple or derived climate variables simulated by the models and the observations are computed, but a final single performance rank is not presented (e.g., Maxino et al, 2008; Nieto S. and Rodríguez-Puebla C., 2006; Russell et al , 2006; Vera et al, 2006, Ulbrich et al, 2008). Other studies propose a single final rank but based on the averaging of skill scores, ranks with or without different relative weight (e.g. Fu et al., 2013, Errasti et al, 2011; 2013). The use of this new single multidimensional index on PDFs would avoid the need to establish a final step in order to combine and weight the different climate variables analyzed, since the multidimensional distributional agreement index S already considers the characteristics of the PDF at the multidimensional space covered by the PDF.

An advantage of our technique is the interpretability of the rankings if the marginal PDFs are used the same way as in Figures 4 to 7. The marginal PDFs can be very easily computed from the 3D PDF (in general the *n*-dimensional PDF) that has been computed to produce the S score. The geometrical structure of the marginal PDFs throws light in terms of the differences between models even when temporal position of every point in the PDF space is lost, since geometrical relations between variables allow us to interpret them. This can not be achieved when using one-dimensional PDFs, since in that case, the relationship of the points between different variables can not be resolved.

4.2 Multivariate kernel density estimators in the context of model performance evaluation methods

In this subsection we consider the methodology for model evaluation based on multidimensional kernel density estimators proposed in this contribution in the general context of the model performance evaluation. Bearing in mind only the two case studies analyzed in the Results section, we could restrict this section to the framework of the modeling of geophysical fluids. There's no reason, however, to restrict the discussion in such a manner. On the contrary, and as already stated, the framework will be that of the evaluation of the performance of environmental models in general. To put the proposed technique in context, we will refer to a recent position paper by Bennett et al. (2013) in this journal on the characterization of the performance of environmental models. The paper by Bennet et al. (2013) describes a series of methods and procedures for both qualitative and quantitative model performance characterization. It also proposes a general purpose five-step recipe for model skill evaluation. Direct references will be made to the classification of evaluation methodologies and the five-step recipe proposed in Bennett at al. (2013), avoiding the inclusion of redundant information or descriptions here. Therefore, the reader is referred to that contribution for a complete understanding of the forthcoming discussion.

In the case of the classification categories of quantitative measures of model performances found in Bennett et al. (2013), the method based on multidimensional kernel density estimators would fall in more than a category at the same time. First, amongst the direct value comparison methods that compare all model and observation values as a whole (even for multidimensional and multivariate cases) to give a single value metric, the multidimensional score that is the result in our case. Note, however, that this classification is valid only for the final step of the technique when the multivariate (multidimensional) PDFs belonging to the observation and the model are compared.

Note also that this PDF to PDF comparison could be understood as a high precision (number of categories) multidimensional contigency table, since PDFs are actually evaluated on a discrete grid. According to this, the last comparison step would fall amongst concurrent comparison methods. In addition, and paying attention to to the step that is carried before the PDFs are compared, it is clear that the method must also be considered within the data transformation method class (transformation to PDF space). Summarizing and according to the classification of quantitative model performance evaluation tools described in Bennett et al. (2013) the proposed technique can be considered as a direct comparison method taken as a whole, but also as a data transformation method and a concurrent comparison method if two inner steps are considered separately.

Although they are not part of the proposed methodology, additional verification techniques considered in Bennet et al. (2013) have also been considered in this contribution, and it is worth mentioning those here. For instance, 2D marginalized PDFs (Figures 4 to 7) were used in the first case study to understand by means of a visual inspection the reasons that drove the relative differences in the scores shown in Table 2. Once again not part of the proposed technique, but mentioned in Bennett et al (2013) and used in our case studies, were the data transformation methods: the Lengendre polynomials or the zonally averaged values in the first case study, and the global-scale PCA analysis of the variables in the second one.

In the case of the five-step general recipe for model performance evaluation proposed by Bennett et al (2013), our procedure should be taken into account in the fifth step of refinements due to its complexity. This does not mean that previous steps like checking the data before any additional step, performing some visual checks and deducing some basic metrics are to be left aside. In fact, they are more than convenient as part of any good practices procedure for model performance evaluation and were also applied in both case studies shown in Results section.

4.3 Additional capabilities and potential applications of the methodology

The case study described in section 3.2 shows the evaluation of a single model with two multidimensional variables. The objective was not to make a thorough verification of the selected CFSR model, but to illustrate some concepts of broad applicability like possibility of using and also combining very different variables (TAS, SST and SSH in the presented case studies), the potential of the use of techniques to reduce the dimensionality of the dataset to whom the verification is to be applied (Legrendre Polinomial of zonally averaged values in the first case study, global-scale principal components in the second case), or the relative score changes of the univariate and

multivariate cases, to mention some.

With regard to the results shown in Table 3, the most remarkable fact is the reduction of the scores in the multivariate case compared to the considerably better scores obtained in the univariate cases. This could be expected, however, as the growing complexity of the PDF with growing number of dimensions/variables will tend to enhance the differences in the model/observation PDFs. This example, although simple, has other potential uses in addition to the one discussed here. For instance, one may want to evaluate the evolution of the performance of a model to see the impacts of developing stages in the model. Then a comparison like the one in case study two will be useful, specially if the improvement can be identified by a multivariate verification. Another application could include the joint evaluation of several model variables, like the one in the first case study. In addition, other variables could be added with selective criteria to the analysis (like the surface currents, the depth of the thermocline or the characterization of ENSO, to mention some) without increasing too much the number of variables, but considering many aspects of the physics accounted by the modelling.

4.4 Limitations of the methodology

As with any other model evaluation index, this index has limitations too. First, it is probably not adequate to detect the probability of very infrequent events, since multidimensional PDF estimations tend to be unreliable on those areas where the value of the PDF is already low. It has been shown in this contribution that its reliability is very low if there exists a bias between models and observations. Secondly, the sample size must be high enough for a modest dimensionality to be analyzed. It is well known that, when analyzing multidimensional datasets, the phase space is too empty when the dimensionality of the space grows (Bellman, 1961). This might also happen with environmental models in general if some kind of initial dimensionality reduction step was not applied to direct model output at the grid point level. Therefore, working with daily data for climatic applications is almost a must. For other kind of environmental models, fast time scales should in any case be used.

Anyway, it has to be considered that multivariate density estimation is always computationally expensive, so that the dimensionality of the dataset must be kept modest. When the dimensionality of the problem increases, the CPU needed increases as a result of the increase of grid points and the complexity of the linear algebra operations performed. Additionally, for high dimensionality, the size of the netCDF files and the memory needed to evaluate them scale too fast and the user is faced

with limitations in the hardware of computers (memory and disk). As a result, we must conclude that it is always necessary to perform an initial step of data dimensionality reduction that allows to work in a fundamental set of low-dimensionality variables that reproduce the behaviour of the system. In the two case studies presented in this contribution, the initial step of dimensionality reduction has been performed by projections onto Legendre polynomials or by means of principal component analysis.

Brands et al. (2012) showed that, in some circumstances, particularly when there exists a clustering of data near zero (such as it happens for specific humidity), the Kolmogorov-Smirnov test can be better applied than the univariate similarity index equivalent to the multidimensional one used in this paper. In addition, the existence of a well known distribution corresponding to the Kolmogorov-Smirnov test allows the researcher to make use of statistical hypothesis testing. This can not be done unless Monte Carlo techniques are used for the univariate similarity index. However, when working in multiple dimensions, the univariate Kolmogorov-Smirnov test can not be extended to several dimensions above two or even three (Fasano and Franceschini, 1987; Justel et al., 1997; Lopes et al., 2008; Peacock, 1983). Since the methodology that we propose in this study can work in multiple dimensions, we find that this methodology will, hopefully, allow to perform an easier evaluation of models according to several criteria, as was originally intended.

5 Conclusions

The index based on the common area between PDFs that is frequently used in the univariate evaluation of climate models, and that is computed as the common area under the PDFs corresponding to models and observations, has been extended to multidimensional problems. In addition, tools that allow its use have been developed and made freely available as open source software. The tools compute the kernel-based multidimensional PDF, identify the optimum value of the bandwidth by means of smoothed bootstrap and compute the common volume under two n-dimensional PDFs.

The use of multidimensional PDFs is very intensive in terms of CPU time, and it is particularly so for the case of the bootstrap estimation of the bandwidth, since several realizations of the PDF must be computed for every bandwidth value tested. The availability of a parallel version of the tool for higher dimensionality and for the bootstrap allows to carry out those computations in short times (less than one hour in our case, using standard hardware).

The use of a multidimensional analysis produces a single index corresponding to every model even after analyzing several variables, and this result makes it easy to perform evaluation of the models under several target variables. In the contribution presented here, we have explored one case such as three Legendre coefficients that expand the daily cycle of zonally averaged temperature. However, the same approach could be applied to the joint analysis of temperature, outgoing longwave radiation or cloud cover (to name a few) such that the structure of the multidimensional PDFs (probably properly marginalized as in this contribution) could shed light over the behaviour of models according to known physical mechanisms.

Even though a tool of the set described in this contribution allows to make an objective selection of the optimal bandwidth to be used in the generation of the PDFs by means of smoothed bootstrap, the case study in this paper shows that the ranking of the models is quite robust even under severe (+/- 30% of the optimal bandwidth) changes in the bandwidth used for the generation of the multidimensional PDFs. Thus, the results obtained through the use of the tools presented in this contribution are reliable.

However, the index is extremely sensitive to the existence of a constant bias between models and it should not be used without previously centering the data, a finding in agreement with previous studies using univariate PDFs (Brands et al., 2011; Brands et al., 2012). The programs do not request that the datasets are centered, but a constant bias between the model and observations could lead to unphysical diagnostics in several dimensions. A potential solution is to perform the evaluation onto *n*-dimensional centered data, so that the bias is automatically removed. Alternatively, the analysis can be performed removing the bias from the model results with respect to observations. Therefore, the analysis of the bias must still be kept independent from the analysis of the shape of the PDF presented in this contribution. The current implementation of the mpdf_score program provides a warning if the bias at any of the dimensions is greater than 5% of the standard deviation of the correspoding variate.

The second case study demonstrated the applicability of the proposed methodology to multivariate and multidimensional data using data from oceanographic SST and SSH variables too. In addition, and although it is not part of the proposed technique, this case study also demonstrated the potential of the use of a preprocessing step for the reduction of the dimensionality of the data, based on a PCA analysis of the original dataset in this case. This shows that the method can potentially be applied to a large family of environmental problems.

The overall evaluation of environmental models is a complex task and different performance scores detect different weak or strong points of the available global models. We hope that the addition of a new methodology and tools that allow its easy application by other researchers make it easier the identification in future experiments of areas of models that can be improved.

Acknowledgements: Authors acknowledge constructive comments by three referees and the editor of this paper. These comments have lead to an improved version of the manuscript. We acknowledge the modeling groups, the Program for Climate Model Diagnosis and Inter-comparison (PCMDI) and the WCRP's Working Group on Coupled Modeling (WGCM) for their roles in making available the WCRP CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, U.S. Department of Energy. ECMWF ERA-40 data used in this study have been provided by ECMWF. *NCEP reanalysis data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at http://www.esrl.noaa.gov/psd/* have been used. CFSR and CFSv2 data was provided by the Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, Colorado. ARMOUR-3D data was obtainned from MyOcean (http://www.myocean.eu/). Authors thank financial funding by project CGL2013-45198-C2-1-R (MINECO, National R+D+i plan), the SAIOTEK program from the Basque Government (project S-P11UN137). Additional funding from different calls from the University of the Basque Country (UFI 11/55, PPM12/01 and GIU 11/01) has allowed this paper to be finished. This work has also been partially supported by the Saiotek and Research Groups 2013-2018 (IT-609-13) programs (Basque Government), TIN2010-14931 (Ministry of Science and Technology), COMBIOMED network in computational bio-medicine (Carlos III Health Institute). U. Lopez-Novoa holds a grant from the Basque Government. J. Miguel-Alonso and A. Mendiburu are members of the HiPEAC European Network of Excellence. Author contributions: JS, AM and JMA designed the research; ULN, JS, AM and JMA wrote the code distributed with this contribution; IE, AE, GIB and JS performed the computations that lead from data in the CMIP3 repository to the Legendre coefficients used in the first case study; GE performed the computations for the second example, ULN, JS, AM and JMA prepared the specific computations, graphics and results used in this contributions after the Legendre coefficients and ULN, JS, AM, IE, GE and JMA wrote the paper.

References

Andrews, T., Gregory, J. M., Webb, M. J., and Taylor, K. E., 2012. Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models. Geophysical Research Letters, 39, L09712, doi: 10.1029/2012GL051607.

Ahamada, I. And Flachaire, E. 2010. Non-parametric Econometrics. Oxford University Press, 176 pages, Oxford.

Annan, J. D., and Hargreaves, J. C. 2010. Reliability of the CMIP3 ensemble, Geophysical Research Letters, 37, L02703, doi:10.1029/2009GL041994.

Bellman, R., 1961. Adaptive Control Processes: A Guided Tour . Princeton University Press, 255 pp.

Bengtsson, L., 2013. What is the climate system able to do “on its own”? Tellus B65, 20189, doi:10.3402/tellusb.v65i0.20189.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.A., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environmental Modelling & Software, 40, 1-20, doi:10.1016/j.envsoft.2012.09.011.

Boer, G. J. and Lambert, S. J., 2001. Second-order space-time climate difference statistics. Climate Dynamics 17, 213-218, doi: 10.1007/PL00013735.

Braconnot, P., Harrison, S. P. , Kageyama, M., Bartlein, P. J. , Masson-Delmotte, V., Abe-Ouchi, A., Otto-Bliesner, B., Zhao, Y., 2012. Evaluation of climate models using palaeoclimatic data , Nature Clim. Change , 2, 417-424, doi: 10.1038/nclimate1456

Brands, S., Herrera, S., San-Martín, D., Gutiérrez, J., 2011. Validation of the ENSEMBLES global climate models over southwestern Europe using probability density functions, from a downscaling perspective. Climate Research, 48, 145-161, doi: 10.3354/cr00995.

Brands, S. Gutiérrez, J.M., Herrera, S., Cofiño, A. S., 2012. On the use of reanalysis data for

downscaling. Journal of Climate 25, 2517-2526, doi: 10.1175/JCLI-D-11-00251.1

Chu, Peter C., 1999. Two Kinds of Predictability in the Lorenz System. Journal of Atmospheric Sciences, 56, 1427–1432. doi: 10.1175/1520-0469(1999)056<1427:TKOPIT>2.0.CO;2

Dagum, L.; Menon, R., 1998 OpenMP: an industry standard API for shared-memory programming, Computational Science & Engineering, IEEE , 5, 46-55, doi: 10.1109/99.660313

DeAngelis, A. M., Broccoli, A. J., Decker, S. G., 2013. A Comparison of CMIP3 simulations of precipitation over North America with observations: Daily statistics and circulation features accompanying extreme events, Journal of Climate, 26, 3209-3230, doi: 10.1175/JCLI-D-12-00374.1

Dessai, S., Lu, X. and Hulme, M., 2005. Limited sensitivity analysis of regional climate change probabilities for the 21st century. Journal of Geophysical research, 110, D19108, doi: 10.1029/2005JD005919.

Dessler, A. E., 2013. Observations of Climate Feedbacks over 2000–10 and Comparisons to Climate Models , Journal of Climate 26, 333-342, doi: 10.1175/JCLI-D-11-00640.1

Duong, T. and Hazelton, M. L., 2005. Cross-validation bandwidth matrices for multivariate kernel density estimation, Scandinavian Journal of Statistics 32, 485-506, doi:10.1111/j.1467-9469.2005.00445.x

Errasti, I., Ezcurra, A., Sáenz, J., Ibarra-Berastegi, G., 2011. Evaluation of IPCC AR4 models over the Iberian Peninsula, Theoretical and Applied Climatology, 103, 61-79, doi: 10.1007/s00704-010-0282-y

Errasti, I., Ezcurra, A., Sáenz, J., Ibarra-Berastegi, G., Zorita, E., 2013. Comparison of the main characteristics of the daily zonally averaged surface air temperature as represented by reanalysis and seven CMIP3 models, Theoretical and Applied Climatology, 114, 417-436, doi: 10.1007/s00704-013-0842-z

Faraway, J. J., Jhun, M., 1990. Bootstrap choice of bandwidth for density estimation. Journal of the American Statistical Association, 85, 1119-1122

Fasano, G., Franceschini, A. 1987. A multidimensional version of the Kolmogorov–Smirnov test, Monthly Notices of the Royal Astronomical Society, 225, 155-170, doi: 10.1093/mnras/225.1.155.

Fernández, J., Primo, C., Cofiño, A. S., Gutiérrez, J. M., Rodríguez, M. A. 2009. MVL spatiotemporal analysis for model intercomparison in EPS: application to the DEMETER multi-model ensemble. Climate Dynamics, 33, 233-243, doi: 10.1007/s00382-008-0456-9

Frederiksen, J. S., Collier, M. A., Watkins, A. B. 2004. Ensemble prediction of blocking regime transitions. Tellus, 56A, 485-500, url: http://www.tellusa.net/index.php/tellusa/article/view/14460.

Fu, G., Liu, Z., Charles, S. P., Xu, Z., Yao, Z., 2013. A score-based method for assessing the performance of GCMs: A case study of southeastern Australia. Journal of Geophysical research, 118, 4145-4167, doi: 10.1002/jgrd.50269

Gates, W. L. , Boyle, J. S., Covey, C., Dease, C. G., Doutriaux, C. M., Drach, R. S., Fiorino, M., Gleckler, P.J., Hnilo, J. J., Marlais, S. M., Phillips, T. J., Potter, G. L., Santer, B. D., Sperber, K. R., Taylor, K. E., Williams, D. N. 1999. An Overview of the Results of the Atmospheric Model Intercomparison Project (AMIP I). Bull. Amer. Meteor. Soc., 80, 29–55,
doi: 10.1175/1520-0477(1999)080<0029:AOOTRO>2.0.CO;2

Gleckler, P. J., Taylor, K. E., Doutriaux, C., 2008. Performance metrics for climate models. Journal of Geophysical Research 113, D22105, doi: 10.1029/2007JD008972

Guinehut S., Le Traon, P.-Y., Larnicol, G., Philipps, S., 2004. Combining Argo and remote-sensing data to estimate the ocean three-dimensional temperature fields - A first approach based on simulated observations. Journal of Marine Systems, 46 (1-4), 85-98, doi: 10.1016/j.jmarsys.2003.11.022

Guinehut S., Dhomps, A.-L., Larnicol, G., Le Traon, P.-Y., 2012. High resolution 3D temperature and salinity fields derived from in situ and satellite observations. Ocean Science, 8(5):845–857, doi:10.5194/os-8-845-2012

Henderson-Sellers, A., Pitman, A. J., Love, P. K., Irannejad, P., Chen, T. 1995. The project for

Intercomparison of land surface parameterisaton schemes PILPS) Phases 2 and 3. Bull. Amer. Meteor. Soc., 76, 489-503, doi: 10.1175/1520-0477(1995)076<0489:TPFIOL>2.0.CO;2

Hewitt, C. D. and Griggs, D. J. 2004. Ensembles-based predictions of climate changes and their impacts, Eos Transactions of the AGU, 85, 566–566, doi:10.1029/2004EO520005.

Hind, A., Moberg, A. Sundberg, R. 2012. Statistical framework for evaluation of climate model simulations by u se of climate proxy data from the last millennium – Part 2: A pseudo-proxy study addressing the amplitude of solar forcing, Climate of the Past, 8 1355-1365,
doi: 10.5194/cp-8-1355-2012.

Justel, A., Peña, D., Zamar, R. 1997. A multivariate Kolmogorov-Smirnov test of goodness of fit. Statistics and Probability Letters, 35, 251-259, doi: 10.1016/S0167-7152(97)00020-5

Kageyama, M., Valdes, P. J., Ramstein, G., Hewitt, C., Wyputta, U. 1999. Northern Hemisphere Storm Tracks in Present Day and Last Glacial Maximum Climate Simulations: A Comparison of the European PMIP Models. Journal of Climate, 12, 742–760,
doi: 10.1175/1520-0442(1999)012<0742:NHSTIP>2.0.CO;2

Kalnay, E. , Kanamitsu, M. , Kistler, R. , Collins, W. , Deaven, D. , Gandin, L. , Iredell, M. , Saha, S. , White, G. , Woollen, J. , Zhu, Y. , Leetmaa, A. , Reynolds, R. , Chelliah, M. , Ebisuzaki, W. , Higgins, W. , Janowiak, J. , Mo, K. C. , Ropelewski, C. , Wang, J. , Jenne, R., Joseph, D., 1996. The NCEP/NCAR 40-year reanalysis project, Bulletin of the American Meteorological Society, 77, 437-470. doi: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2

Knutti, R., 2008. Should we believe model predictions of future climate change? Philosophical Transactions of the Royal Society A, 366, 4647-4664, doi: 10.1098/rsta.2008.0169.

Knutti, R., 2010. The end of model democracy? An editorial comment, Climatic Change 102, 395-404, doi: 10.1007/s10584-010-9800-2

Knutti, R., Masson, D., Gettelman, A., 2013. Climate model genealogy: Generation CMIP5 and how we got there. Geophysical Research Letters, 40, 1194-1199, doi:10.1002/grl.50256.

Lamarque, J.-F., Shindell, D. T., Josse, B., Young, P. J., Cionni, I., Eyring, V., Bergmann, D., Cameron-Smith, P., Collins, W. J., Doherty, R., Dalsoren, S., Faluvegi, G., Folberth, G., Ghan, S. J., Horowitz, L. W., Lee, Y. H., MacKenzie, I. A., Nagashima, T., Naik, V., Plummer, D., Righi, M., Rumbold, S. T., Schulz, M., Skeie, R. B., Stevenson, D. S., Strode, S., Sudo, K., Szopa, S., Voulgarakis, A., and Zeng, G. 2013. The Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP): overview and description of models, simulations and climate diagnostics, Geosci. Model Dev., 6, 179-206, doi:10.5194/gmd-6-179-2013.

Lopes, R. H. C., Hobson, P. R., Reid, I. D. 2008. Computationally efficient algorithms for the two-dimensional Kolmogorov–Smirnov test. Journal of Physics: Conference Series, 119, 042019, doi:10.1088/1742-6596/119/4/0420

Lorenz, E. N. 2006. Predictability, a problem partly solved, Chapter 3 in Palmer, T. and Hagedorn, R. (eds.) Predictability of Weather and Climate, Cambridge University Press, Cambridge, 702 pp.

Macadam, I., Pitman, A. J., Whetton, P. H., Abramowitz, G., 2010. Ranking climate models by performance using actual values and anomalies: Implications for climate change impact assessments. Geophysical Research Letters 37, L16704, doi: 10.1029/2010GL043877.

Marshall, A.G., Hudson, D., Hendon, H.H., Pook, M. Alves,O., Wheeler, M. 2013. Simulation and prediction of blocking in the Australian region and its influence on intra-seasonal rainfall in POAMA-2. Climate Dynamics doi: 10.1007/s00382-013-1974-7.

Masson, D., Knutti, R., 2011. Climate model genealogy, Geophysical Research Letters, 38, L08703, doi: 10.1029/2011GL046864.

Maxino, C. C., McAvaney, B. J., Pitman, A. J., Perkins, S. E., 2008. Ranking the AR4 climate models over the Murray-Darling Basin using simulated maximum temperature, minimum temperature and precipitation. International Journal of Climatology, 28, 1097-1112, doi: 10.1002/joc.1612.

Meehl, G. A., Covey, C., Taylor, K. E., Delworth, T., Stouffer, R. J., Latif, M., McAvaney, B., Mitchell, J. F. B., 2007. The WCRP CMIP3 multimodel dataset: A new era in climate change research. Bulletin of the American Meteorological Society, 88, 1383-1394. doi: 10.1175/BAMS-

88-9-1383

Menke, W. and Menke, J., 2012, Environmental Data Analysis with Matlab, 263 pages, Elsevier, Oxford.

Moberg, A. 2013. Comparisons of simulated and observed Northern Hemisphere temperature variations during the past millennium – selected lessons learned and problems encountered: Tellus B 65, 19921, doi: 10.3402/tellusb.v65i0.19921.

Nieto S. and Rodríguez-Puebla C. 2006. Comparison of Precipitation from observed data and general circulation models over the Iberian Peninsula. Journal of Climate, 19: 4254-4275.

North, G. R., Cahalan, R. F., Coakley, J. A., 1981. Energy balance climate models. Review of Geophysica and Space Physics, 19:91-121, doi: 10.1029/RG019i001p00091

Otto, A., Otto, F., Boucher, O., Church, J., Hegerl, G., Forster, P. M., Gillett, N. P., Gregory, J., Johnson, G. C., Knutti, R., Lewis, N., Lohmann, U., Marotzke, J., Myhre, G., Shindell, D., Stevens, B., Allen, M.R. 2013. Energy budget constraints on climate response, Nature Geoscience, 6, 415–416, doi:10.1038/ngeo1836

Peacock, J. A. 1983. Two-dimensional goodness-of-fit testing in astronomy, Monthly Notices of the Royal Astronomical Society, 202, 615-627, doi: 10.1093/mnras/202.3.615.

Pelly, J. L., Hoskins, B. J. 2003. How well does the ECMWF Ensemble Prediction System predict blocking?. Quarterly Journal of the Royal Meteorological Society, 129, 1683–1702. doi: 10.1256/qj.01.173

Pennell, C., Reichler, T., 2011. On the Effective Number of Climate Models. Journal of Climate, 24, 2358–2367, doi: 10.1175/2010JCLI3814.1

Perkins, S. E., Pitman, A. J., N. H. Holbrook, McAneney, J., 2007. Evaluation of the AR4 climate models' simulated maximum temperature, minimum temperature and precipitation over Australia using probability density functions. Journal of Climate 20, 4356-4376, doi: 10.1175/JCLI4253.1

Radić, R., Clarke, G. K., 2011. Evaluation of IPCC Model's performance in simulating Late-Twentieth-Century climatologies and weather patterns over North America. Journal of Climate 24, 5257-5274, doi: 10.1175/JCLI-D-11-00011.1

Randall, D. A., Wood, R. A., Bony, S., Colman, R. Fichefet, F., Fyfe, J., Kattsov, V., Pitman, A., Shukla, J., Srinivasan, J., Stouffer, R. J., Sumi, A., Taylor, K. E., 2007. Climate models and their evaluation. In: Solomon, S., Qin, D., Manning, Chen, Z., M., Marquis, M., Averyt, K., Tignor, M. M. B., Miller, H. L., Climate Change 2007, The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge , UK and New York, NY, USA, 996pp.

Reichler, T. and Kim, J. K., 2008. How well do coupled models simulate today's climate? Bulletin of the American Meteorological Society 89, 303-311, doi: 10.1175/BAMS-89-3-303

Reifen, C. and Toumi, R., 2009. Climate projections: past performance no guarantee of future skill? Geophysical Research Letters, 36, L13704, doi: 10.1029/2009GL038082.

Russell J., Stouffer R.J. and Dixon K.W. 2006. Intercomparison of the Southern Ocean circulations in IPCC coupled model control simulations. Journal of Climate, 19: 4560–4575.

Saha, S., Moorthi, S., Pan, H., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y., Chuang, H., Juang, H. H., Sela, J., Iredell, M., Treadon, R., Kleist, D., Van Delst, P., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., Van Den Dool, H., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R. W., Rutledge, G., Goldberg, M., 2010. The NCEP Climate Forecast System Reanalysis. Bulletion of the American Meteoological Society, 91, 1015–1057, doi: 10.1175/2010BAMS3001.1

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y., Chuang, H., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez Malaquías, P., van den Dool, H., Zhang, Q., Wang, **W., Chen, M., Becker, E., 2014. The NCEP Climate Forecast System Version 2. Journal of Climate, 27, 2185–2208. doi: 10.1175/JCLI-D-12-00823.1**

Santer, B. D., Mears, C., Doutriaux, C., Caldwell, P., Gleckler, P. J., Wigley, T. M. L., Solomon, S., Gillet, N. P., Ivanova, D., Karl, T. R., Lanzante, J. R., Meehl, G. A., Stott, P. A., Taylor, K. E., Thorne, P. W., Wehner, M. F., Wentz, F. J., 2011. Separating signal and noise in atmospheric temperature changes: The importance of timescale. Journal of Geophysical research, 116, D22105, doi: 10.1029/2011JD016263.

Schwalm, C. R., Huntinzger, D. N., Michalak, A. M., Fisher, J. B., Kimball, J. S., Mueller, B., Zhang, K. and Zhang, Y., 2013. Sensitivity of inferred climate model skill to evaluation decisions: a case study using CMIP5 evapotranspiration. Environmental Research Letters, 8, 024028, doi: 10.1088/1748-9326/8/2/024028

Scott, D. W. 1992. Multivariate Density Estimation: Theory, Practice and Visualization, John Wiley and Sons, New York, 336 pp, doi: 10.1002/9780470316849

Silverman, B. W., 1986. Density Estimation for Statistics and data Analysis, Chapman and Hall, London, 175pp.

Stewart, D. E. and Leyk, Z., 1994. Meschach: Matrix Computations in C, Centre for Mathematics and its Applications, the Australian National University, Canberra, Australia.

Stocker, T.F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P. M. (eds.). 2013. IPCC, 2013: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp.

Sundberg, R., Moberg, A., and Hind, A. 2012. Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium – Part 1: Theory, Climate of the Past, 8, 1339-1353, doi: 10.5194/cp-8-1339-2012.

Taylor, K. E., 2001. Summarizing multiple aspects of model performance in a single diagram, *Journal of Geophysical Research, 106, 7183–7192, doi: 10.1029/2000JD900719.*

Taylor, K. E., Stouffer, R. J., Meehl, G. A., 2012. An overview of CMIP5 and the experiment

design. Bulletin of the American Meteorological Society, 93, 485-498, doi: 10.1175/BAMS-D-11-00094.1

Ulbrich U, Pinto JG, Kupfer H, Leckebusch G, Spangehl T and Reyers M. (2008). Changing Northern Hemisphere Storm Tracks in an Ensemble of IPCC Climate Change Simulations. Journal of Climate, 21,1669-1679.

Uppala, S.M., Kållberg, P.W., Simmons, A.J., Andrae, U., da Costa Bechtold, V., Fiorino, M., Gibson, J.K., Haseler, J., Hernandez, A., Kelly, G.A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R.P., Andersson, E., Arpe, K., Balmaseda, M.A., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B.J., Isaksen, L., Janssen, P.A.E.M., Jenne, R., McNally, A.P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N.A., Saunders, R.W., Simon, P., Sterl, A., Trenberth, K.E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J. 2005. The ERA-40 re-analysis. Quarterly Journal Royal Meteorological Society, 131, 2961-3012. doi: 10.1256/qj.04.176

Vera C., Silvestri G., Liebmann B. and Gonzalez P., 2006. Precipitation variability in South America from IPCC-AR4 models. Part II: influence of circulation leading patterns. Proceedings of 8 ICSHMO, Foz do Iguaçu, Brazil, INPE, April 24–28: 477–485.

von Storch, H., Zwiers, F. 1999. Statistical Analysis in Climate Research, Cambridge University Press, Cambridge, 484 pp.

Walsh, J. E., Chapman, W. L., Romanovsky, V., Christensen, J. H., Stendel, M., 2008. Global climate model performance over Alaska and Greenland. Journal of Climate 21, 6156-6174, doi: 10.1175/2008JCLI2163.1

Wang, M., Overland, J. E., Kattsov, V., Walsh, J. E., Zhang, X., Pavlova, T. 2007. Intrinsic versus Forced Variation in Coupled Climate Model Simulations over the Arctic during the Twentieth Century. Journal of Climate, 20, 1093–1107. doi: 10.1175/JCLI4043.1

Ylhäisi, J. S. and Räisänen, J., 2013. Twenty-first century changes in daily temperature variability in CMIP3 climate models. International Journal of Climatology, In Press, doi: 10.1002/joc.3773.

Washington, W. M., Parkinson, C. L., 2005, An Introduction to Three-Dimensional Climate Modelling, 2nd ed. University Science Books, 353 pages, Sausalito, CA.

Wilks, D. S., 2006. Statistical Methods in the Atmospheric Sciences, 2nd ed. Academic Press, Burlington, 627 pp.

Table 2. Values of the multidimensional ***S*** score corresponding to different values of the bandwidth parameter and associated rankings that would correspond to the models, when compared with ERA40 reanalysis data.

	h = 0.4		h = 0.5		h = 0.6		h = 0.7		h = 0.8	
Model	S score	rank	S score	rank	S score	rank	S score	rank	S score	rank
BCM2.0	0.45	10	0.48	9	0.51	9	0.53	9	0.56	9
ECHAM5	0.45	9	0.47	10	0.48	10	0.50	10	0.51	10
GFDL-CM2.0	0.55	8	0.58	8	0.60	8	0.61	8	0.63	8
GFDL-CM2.1	0.58	7	0.60	7	0.62	7	0.64	7	0.65	7
HADGEM1	0.66	5	0.69	4	0.71	4	0.72	4	0.73	4
MIROCS3.2-HR	0.62	6	0.65	6	0.67	6	0.70	6	0.71	6
MIROCS3.2-MR-RUN01	0.66	4	0.68	5	0.70	5	0.72	5	0.73	5
MIROCS3.2-MR-RUN02	0.69	2	0.72	2	0.74	2	0.75	2	0.77	2
MIROCS3.2-MR-RUN03	0.69	3	0.71	3	0.73	3	0.74	3	0.75	3
MRI-RUN01	0.25	13	0.27	13	0.29	13	0.31	13	0.33	14
MRI-RUN02	0.25	14	0.27	14	0.29	14	0.31	14	0.33	13
MRI-RUN03	0.25	11	0.28	11	0.30	11	0.32	11	0.34	11
MRI-RUN04	0.25	12	0.27	12	0.29	12	0.32	12	0.34	12
MRI-RUN05	0.23	15	0.25	15	0.28	15	0.30	15	0.32	15
NCEP	0.80	1	0.81	1	0.82	1	0.83	1	0.84	1

Table 4. Global rankings for the models according to their representation of the daily zonally averaged temperature in Errasti et al. (2013) and according to the methodology proposed in this contribution (3D-PDF).

	Errasti et al. 2013 rank	3D-PDF rank
BCM2.0	5	6
GFDL-CM2.0	7	5
GFDL-CM2.1	6	4
MIROCS3.2-HR	2	3
MIROCS3.2-MR	3	2
ECHAM5	4	7
MRI-CGCM2.3	8	8
NCEP	1	1

Multi-objective environmental model evaluation by means of multidimensional kernel density estimators: efficient and multi-core implementations

Highlights:

- The performance index based on the area under two PDFs is extended to several dimensions.
- The evaluation of the performance of models can be done for several variables, resulting in a single skill score.
- A fast and parallel implementation that allows to apply the method with highly dimensional problems is presented.
- The method is illustrated with two case-studies.
- The sensitivity of the results to the bias between models and observations or the bandwidth is presented.