



**HAL**  
open science

## Web processing service for climate impact and extreme weather event analyses. Flyingpigeon (Version 1.0)

Nils Hempelmann, Carsten Ehbrecht, M Carmen Alvarez-Castro, Patrick Brockmann, Wolfgang Falk, Jörg Hoffmann, Stephan Kindermann, Ben Koziol, Cathy Nangini, Sabine Radanovics, et al.

### ► To cite this version:

Nils Hempelmann, Carsten Ehbrecht, M Carmen Alvarez-Castro, Patrick Brockmann, Wolfgang Falk, et al.. Web processing service for climate impact and extreme weather event analyses. Flyingpigeon (Version 1.0). *Computers & Geosciences*, 2018, 110, pp.65-72. 10.1016/j.cageo.2017.10.004 . hal-01375615v2

**HAL Id: hal-01375615**

**<https://hal.science/hal-01375615v2>**

Submitted on 21 May 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Web processing service for climate impact and extreme weather event analyses. Flyingpigeon (Version 1.0)

Nils Hempelmann<sup>a,\*</sup>, Carsten Ehbrecht<sup>b,\*</sup>, Carmen Alvarez-Castro<sup>a</sup>, Patrick Brockmann<sup>a</sup>, Wolfgang Falk<sup>c</sup>, Jörg Hoffmann<sup>d</sup>, Stephan Kindermann<sup>b</sup>, Ben Koziol<sup>e</sup>, Cathy Nangini<sup>a</sup>, Sabine Radanovics<sup>a</sup>, Robert Vautard<sup>a</sup>, Pascal Yiou<sup>a</sup>

<sup>a</sup>*Le Laboratoire des Sciences du Climat et de l'Environnement*

<sup>b</sup>*German Climate Computing Center*

<sup>c</sup>*Bayerische Landesanstalt für Wald und Forstwirtschaft*

<sup>d</sup>*Julius Kühn-Institut - Federal Research Centre for Cultivated Plants*

<sup>e</sup>*NOAA Environmental Software Infrastructure and Interoperability Group/University of Colorado-Boulder*

---

## Abstract

Analyses of extreme weather events and their impacts often requires big data processing of ensembles of climate model simulations. Researchers generally proceed by downloading the data from the providers and processing the data files “at home” with their own analysis processes. However, the growing amount of available climate model and observation data makes this procedure quite awkward. In addition, data processing knowledge is kept local, instead of being consolidated into a common resource of reusable code. These drawbacks can be mitigated by using a web processing service (WPS). A WPS hosts services such as data analysis processes that are accessible over the web, and can be installed close to the data archives.

We developed a WPS named ‘flyingpigeon’ that communicates over an HTTP network protocol based on standards defined by the Open Geospatial Consortium (OGC) [23], to be used by climatologists and impact modelers as a tool for analyzing large datasets remotely.

---

\*Corresponding author

*Email addresses:* [info@nilshempelmann.de](mailto:info@nilshempelmann.de) (Nils Hempelmann),  
[ehbrecht@dkrz.de](mailto:ehbrecht@dkrz.de) (Carsten Ehbrecht)

Here, we present the current processes we developed in flyingpigeon relating to commonly-used processes (preprocessing steps, spatial subsets at continent, country or region level, and climate indices) as well as methods for specific climate data analysis (weather regimes, analogues of circulation, segetal flora distribution, and species distribution models). We also developed a novel, browser-based interactive data visualization for circulation analogues, illustrating the flexibility of WPS in designing custom outputs.

Bringing the software to the data instead of transferring the data to the code is becoming increasingly necessary, especially with the upcoming massive climate datasets.

*Keywords:* Web Processing Service, climate impact, extreme weather events, birdhouse, OGC

---

## 1. Introduction

Processing of climate data is typically carried out by individual researchers, who create and run their own scripts in their preferred programming language, either locally or in environments internal to their institutions. Thus, there is a vast but unconnected body of knowledge that is not readily available to the climate science community with the risk of being continually replicated as researchers write scripts for processes that have already been well-developed by others. Furthermore, climate data, such as the upcoming Phase 6 of the Coupled Model Intercomparison Project [CMIP6](#) and the Coordinated Regional Climate Downscaling Experiment [CORDEX](#), is becoming too large to download and process locally.

Here, we present a web processing service (WPS) named 'flyingpigeon' (Version 1.0) containing processes written for and by climatologists and impact modelers for climate impact and extreme weather events analyses. These users are experts in their scientific fields with a good knowledge of climate model data usage, including the uncertainties associated with the data and the methods implemented in the processes.

Flyingpigeon (henceforth always referring to Version 1.0) is part of the

21 open source project [birdhouse](#) (under the Apache License 2.0), a collection of  
22 Open Geospatial Consortium (OGC) WPSs that provides data processing for  
23 the climate science community. Like all other compartments in birdhouse,  
24 flyingpigeon communicates over the web using the HTTP protocol based  
25 WPS Interface Standard for geospatial processing services defined by OGC  
26 [23]. A make file that handles appropriate software dependencies makes it  
27 easy to install. Flyingpigeon and the processes we developed are freely avail-  
28 able from the [flyingpigeon repository](#) on GitHub. To run efficiently, the code  
29 should be installed on a system with appropriate resources.  
30 Birdhouse evolved out of data management projects (C3Grid-INAD [19],  
31 ExArch [3]), LSDMA [17]), while the processes designed for flyingpigeon were  
32 developed within projects related to climate impact and extreme weather  
33 events analysis ([EUCLEIA](#), [A2C2](#), [Extremoscope](#)).  
34 Inter-WPS communication enables operative services for international collab-  
35 orations such as the Infrastructure for the European Network of Earth System  
36 Modeling ([IS-ENES](#)) and Earth System Grid Federation ([ESGF](#)), and is in  
37 line with other WPS developments like [52° North](#) [2] (enabling standardized  
38 deployment of geo-processes on the web), the [ZOO-Project](#) [10](able to pro-  
39 cess geospatial or non geospatial data online), [climate4impact](#), and Climate  
40 Information Portal of Coperincus ([CLIPC](#) [7]).  
41 The goal of this paper is to introduce flyingpigeon as a WPS for climatologists  
42 and impact modelers.

## 43 2. WPS general description

44 A WPS is a technical solution ([WPS Concepts](#)) in which processes are  
45 hosted on a server and accessed over the web (Fig. 1). These processes con-  
46 form to a standardized format, ensuring that they follow the principle of  
47 reusable design: they can be instantiated multiple times for different input  
48 arguments or data sources, customized following the same structure to handle  
49 new inputs, and are modular, hence can be combined to form new processes.  
50 In addition, a WPS can be installed close to the data to enable processing  
51 directly out of the archive. A WPS can also be linked to a theoretically limit-  
52 less combination of several other WPSs, or generally OpenGIS Web Services  
53 ([OWS](#)).

54 In this paper **process** is used in the same sense as in the OGC standard: *'for*  
55 *any algorithm, calculation or model that either generates new data or trans-*  
56 *forms some input data into output data'* [23]. A submitted process is a **job**.

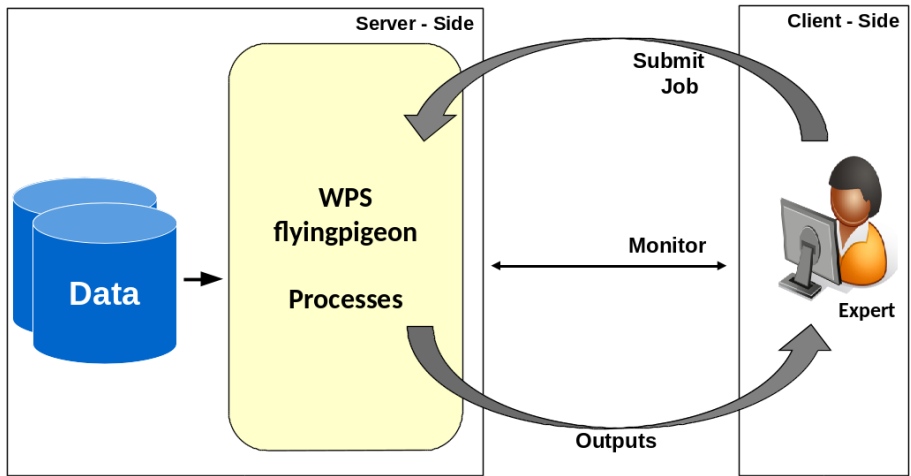


Figure 1: Schematic of WPS operations.

57 A **service** provides a collection of processes containing scientific **methods**  
 58 that focus on climate impact and extreme weather events. A combination  
 59 of processes is called a **workflow**, and a collection of WPS-related software  
 60 compartments is a **framework** (see Section 2).  
 61 WPS divides the operation into **server** and **client** side 2.3, with appropriate  
 62 **security** 2.4 in between to avoid misuse.

### 64 2.1. Birdhouse

65 Birdhouse is a collection of WPS-related Python components to support  
 66 data processing in the climate science community according to their own  
 67 needs and use cases. In birdhouse, we currently use the Python implemen-  
 68 tation of WPS, [PyWPS](#), but birdhouse is not restricted to a single WPS  
 69 implementation. Birdhouse is not “yet another” processing framework, in-  
 70 stead it provides the “glue” and the missing parts to successfully run WPS  
 71 for climate data processing.

72 Birdhouse consists of several components like Flyingpigeon and Emu (see  
 73 Fig. 2). Each of them can be installed individually. The installation is  
 74 done using the Python-based build system [Buildout](#) and [Ansible](#). Most of  
 75 the dependencies are maintained in the [Anaconda](#) Python distribution. For  
 76 convenience, each birdhouse component has a Makefile to ease the installation  
 77 so you don’t need to know how to call the build tools.

78 For managing and interacting with processing services, Birdhouse uni-  
79 formly exposes OGC WPS standard based interfaces. The OGC WPS inter-  
80 face descriptions can be registered in an OGC Web Catalog Service support-  
81 ing standards-based service discovery. Processing results can be published in  
82 the same Catalog Service.

83 Birdhouse has a web-client “Phoenix” to interact with web processing  
84 services and to feed them with data from climate data archives.

85 To control the user access to WPS services (and other OGC services),  
86 birdhouse has an OGC Web Service (OWS) security proxy “Twitcher” which  
87 can be placed in front of any WPS service.

88 Birdhouse has several web processing services which combine processes  
89 related to different aspects of climate data processing. Currently these are:

- 90 • **Flyingpigeon** contains a variety of processes ranging from simple poly-  
91 gon subsetting to complex data analysis methods and workflows used  
92 in climate impact or extreme weather event studies. Flyingpigeon is  
93 the main focus of this paper.
- 94 • **Hummingbird** provides processes to check conformance to climate  
95 metadata standards. These standards are the [NetCDF-CF](#) (Climate  
96 and Forecast conventions) and metadata conventions of climate data  
97 simulation projects like CORDEX and CMIP6.
- 98 • **Malleefowl** has processes to access climate data archives like the Earth  
99 System Grid Federation (ESGF) and Thredds data catalogs. It includes  
100 a workflow process to fetch climate data from a selected archive and  
101 provides this data to a selected analysis process. If the requested cli-  
102 mate data files are not already locally available on disk, they will be  
103 downloaded and cached on the file-system.
- 104 • **Emu** has some lightweight processes to show which input and output  
105 parameters are supported by WPS and provides examples for writing  
106 your own processes.

## 107 2.2. WPS *flyingpigeon*

108 Flyingpigeon is based on [pywps4](#) and contains a variety of processes rang-  
109 ing from simple polygon subsetting to complex data analysis methods and  
110 workflows used in climate impact or extreme weather event studies (see Sec-  
111 tion 3).

112  
113 One of the main software components of flyingpigeon is [OpenClimateGIS](#)  
114 (OCGIS), an open source Python package designed for geospatial manipu-

115 lation, subsetting, computation, and translation of climate datasets stored  
116 locally in NetCDF files or served via OPeNDAP protocols. OCGIS inter-  
117 prets the climate data community’s canonical metadata standard – the [Cli-](#)  
118 [mate and Forecast \(CF\) Convention](#) – maintaining standards compliance for  
119 any derived CF-NetCDF output. OCGIS also supports numerous time-aware  
120 computations (i.e. monthly mean, seasonal maximum), multiple output for-  
121 mats (i.e. ESRI Shapefile, Comma Separated Value), and spatial interpola-  
122 tion using the [Earth System Modeling Framework](#) (ESMF). OCGIS calls can  
123 be executed on the whole dataset at once or divided into chunks to reduce  
124 the memory load. In flyingpigeon, memory availability is checked before and  
125 the call is executed accordingly.

126 Flyingpigeon is mainly written in Python but also includes code parts based  
127 on [CDO](#) commands, R scripts or Fortran and are called out of the python  
128 code. The birds hummingbird, malleefowl and emu currently implemented in  
129 birdhouse are also based on python but with respect of the OGC standard  
130 it is also possible to include WPS services (other birds) completely written in  
131 other languages than python.

132 After submitting a job to the WPS, several data preprocessing steps are run  
133 automatically. Flyingpigeon has utilities to e.g. sort files belonging to differ-  
134 ent datasets, checks the variable contained in the data files, converts units,  
135 rotates or unrotates grid coordinates, or generates file names based on the  
136 metadata or its values.

137 An installation system simplifies the deployment of flyingpigeon on a server.  
138 It is designed for Linux distributions and fetches required dependencies with  
139 the [Conda](#) package management system. It uses [Buildout](#) to setup the ap-  
140 plication by calling a simple install command.

141 Since the code is open source, contributions by the community are possi-  
142 ble. [Online documentation](#) is generated with Sphinx [1], including [automatic](#)  
143 [documentation](#) for the functions.

### 144 *2.3. Server - Client side*

145 Flyingpigeon is a server-side service embedded in the birdhouse (see Fig. 2),  
146 a framework to support the development and deployment of climate data  
147 processing services based on the OGC WPS standard. Flyingpigeon can be  
148 run in combination with other birdhouse compartments such as **phoenix**,  
149 a graphical user interface (GUI) for common web browsers that includes an  
150 OGC web mapping service (WMS), or **malleefowl**, the WPS server backend  
151 for data search in and fetching from the ESGF archive. Other compart-

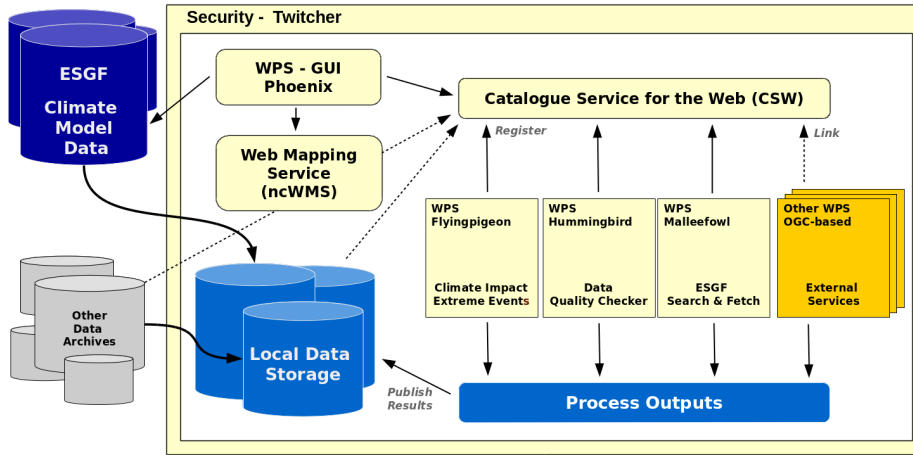


Figure 2: Main components of the birdhouse framework and connection to data archives.

152 ments offer processes such as quality checks for technical aspects of NetCDF  
 153 files ([climate forecast](#) compliance checker) or processes to execute climate  
 154 data operations with CDO. A catalog service for the web ([CSW](#)) is used to  
 155 publish results and discover available services. However, flyingpigeon is a  
 156 self-contained service and can also be run operationally as a stand-alone.

157

158 On the client side, the user needs to connect to the server with an ap-  
 159 propriate URL to enable communication via an HTTP protocol, which is  
 160 possible in three ways:

- 161 • with a terminal command
- 162 • within a script language (e.g. Python)
- 163 • with a browser-based GUI (e.g. Phoenix)

164 Besides connecting to the server via a terminal or a script language, a job  
 165 can be submitted via the Phoenix GUI, directly or through a user-friendly  
 166 wizard that guides through the steps from data search to literal argument  
 167 input, and finally monitors the status of the job.

168

#### 169 *2.4. Security*

170 Web Processing Services are HTTP services and access to them can be  
 171 secured by using HTTPS and firewalls. One can enable basic authentication  
 172 on the HTTPS service by requiring a username and password to use the WPS



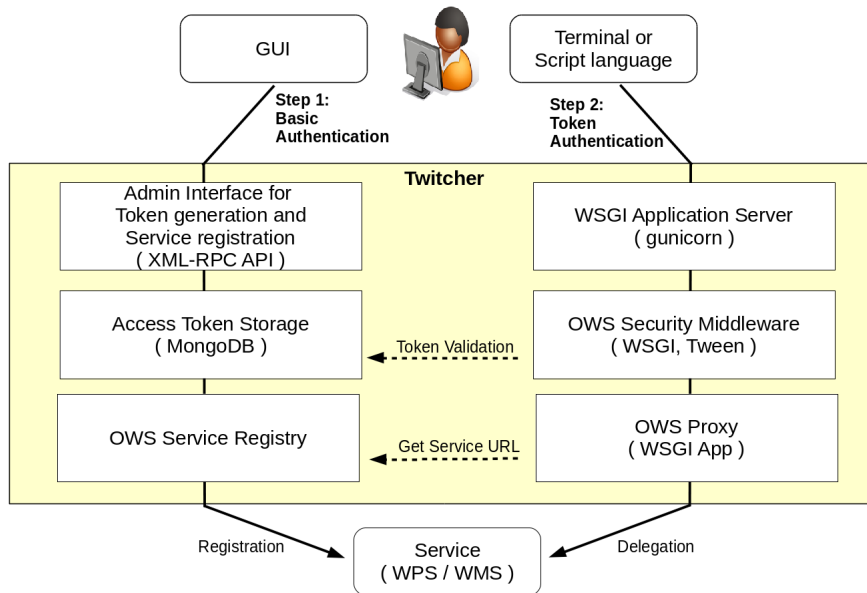


Figure 3: Twitcher security schema.

173 service. Most WPSs are used in an internal network and their processing ca-  
 174 pabilities are used by web portals. External users usually don't have direct  
 175 access to the WPS itself.

176 Birdhouse uses [Twitcher](#), a security proxy we developed for WPSs that pro-  
 177 vides an easy-to-use mechanism to access WPS directly by users in a secure  
 178 way. The concept is not restricted to WPS and can be extended to other  
 179 services like Web Mapping Services (WMS).

180 The Twitcher security proxy protects registered WPSs and allows the execu-  
 181 tion of processes over a terminal or within a script only if a valid access token  
 182 is provided. A basic authentication via a GUI with username and password  
 183 (Step 1 in Fig. 3) allows access to the Twitcher administration interface and  
 184 triggers a token (unique string) generation. This token can be used to access  
 185 the registered WPS services on the command line or in a script (Step 2 in  
 186 Fig. 3). Tokens are valid only for a short period of time and can be regener-  
 187 ated by repeating Step 1. Without a valid token, it is not possible to execute  
 188 a process, but it is still possible to retrieve information about the service to  
 189 explore provided processes and their descriptions.

190

191 The management of the access tokens and the registered WPS is con-  
192 trolled by the Twitcher administration interface using the [XML-RPC](#) proto-  
193 col (first layer in Fig. 3, left column). Access tokens are generated, validated  
194 and persisted by the Access Token Storage (second layer) which uses a [Mon-](#)  
195 [goDB](#) database. The OWS Service Registry (third layer) stores the registered  
196 WPS services with the service URL and a unique service name.

197 In Step 2, the Twitcher proxy service uses a [WSGI](#) application service  
198 (first layer in Fig. 3, right column), a specification for communication between  
199 web servers and web applications. The OWS Security Middleware (second  
200 layer) gets the provided token from the request (generated in Step 1 and  
201 embedded in the URL or an HTTP header variable) and validates it using  
202 the Access Token Storage. If the token is valid, the OWS Proxy (third layer)  
203 retrieves the service URL for the given service name and delegates the request  
204 to the registered service (WPS, WMS, etc.).

## 205 2.5. Input Data

206 Climate model data are commonly stored in NetCDF file format and or-  
207 ganized in data archives. If data are not already stored on the WPS server,  
208 they have to be fetched and stored in a structured way. Besides many other  
209 datasets, the output of the CMIP5 (and upcoming CMIP6) and CORDEX  
210 are stored for public access in the ESGF data archive. Birdhouse provides a  
211 data search for ESGF as a search interface within the GUI or from the com-  
212 mand line. A second important group of datasets are reanalyses data, which  
213 are outputs from a data assimilation component of a weather forecast model.  
214 In climate impact and extreme event assessment, the use of reanalyses data  
215 is very common, thus some processes (see Section 3) in flyingpigeon provide a  
216 preselection of reanalyses datasets and variables. By selecting an appropriate  
217 variable, the data are automatically fetched within the process.

218 In the processes for analogue circulation (3.4.2) and weather regimes (3.4.1),  
219 the preselection of variables are dataset subsets of the following global re-  
220 analysis projects:

- 221 • *NCEP/NCAR Reanalysis 1:*  
222 NCEP data [18] are available from 1948 to the present in 17 pressure  
223 levels with a spatial grid resolution of  $2.5^\circ \times 2.5^\circ$
- 224 • *20th Century Reanalysis version 2 (20CR):*  
225 The 20CR dataset [5] is based on surface pressure observations only and  
226 provides an ensemble of 56 members that contains global weather con-  
227 ditions and their uncertainty from 1871–2012, available on 24 pressure

228 levels with a spatial grid resolution of 2° x 2°.  
229 Local data stored on the server side can also be directly provided as input  
230 data for these two processes.  
231 In climate impact and extreme event assessments, some analytic methods  
232 require non-climatic data. The species distribution model (SDM) process  
233 (3.3.2) in flyingpigeon to predict favorability of tree species requires geograph-  
234 ical coordinates of tree occurrences. A free and open access data base for  
235 biodiversity data is provided by the Global Biodiversity Information Facil-  
236 ity (GBIF), where observations of trees (and other species) are stored and  
237 automatically fetched by the SDM process.

### 238 3. Processes embedded in flyingpigeon

239 Flyingpigeon is a service targeting climate researchers and experts in gen-  
240 eral with a focus on climate impact models and extreme event analyses. This  
241 section describes the processes that we developed so far.

242

#### 243 3.1. Extracting spatial subsets

244 A standard operation in climate analysis is the extraction of data from  
245 a specific spatial region, a process known as polygon spatial subsetting. In  
246 flyingpigeon, the polygon subset process is possible at three levels of increas-  
247 ing spatial resolution: continents, world countries and European administra-  
248 tive regions obtained from the global administrative areas GADM database  
249 (*gadm26\_levels.gdb, v2.5*). To optimize subsetting performance, high reso-  
250 lution GADM boundaries were simplified [42] using [mapshaper](#) [12] with a  
251 1% point retention. The coarse resolution of target climate model inputs  
252 compared with the down-sampled resolution of GADM boundaries did not  
253 result in any significant data loss following a spatial subset operation. Very  
254 small administrative regions (e.g. cantons of Switzerland) were merged on  
255 the country level. For further details, please see the [online documentation](#).  
256 If more than one polygon is selected, a sperate netCDF file will be provided  
257 for each selected dataset. However, if the mosaic option is checked, the se-  
258 lected polygons are merged into one polygon. This increases the flexibility  
259 to meet specific user needs. The provided NetCDF files are wrapped in a tar  
260 archive.

261 In addition to polygons, a process is also available to extract longitude and  
262 latitude points. These point inspection data products are returned in text

263 files containing a timeseries table for each longitude and latitude point.

264

### 265 *3.2. Computation of climate indices*

266 Climate indices are metrics to describe climate conditions and can be  
267 used to assess changes in climate over time (see the use case below), or as  
268 forcing data for complex impact models (e.g. the species distribution model  
269 [3.3.2](#)). Climate indices can be user-defined, but there is also a standardized  
270 set provided by the [European Climate Assessment & Dataset](#) website.

271

272 The climate indices are calculated based on a time aggregation (e.g. year,  
273 month, season) and include world countries polygon subsetting as an optional  
274 feature. The processes fall into four categories (the last two are still under  
275 development):

276

- 277 1. standard
- 278 2. percentile-based
- 279 3. multi-variable indices
- 280 4. user-defined

281 Standard indices are based on one daily input variable (near surface tem-  
282 perature, precipitation). The provided indices are related to the [ICCLIM](#)  
283 Python package deployed in OCGIS. In the percentile-based process, a per-  
284 centile is calculated over a user defined reference period. The process outputs  
285 the number of days that exceed or fall below this calculated threshold (for  
286 percentiles  $> 50$  or  $< 50$ , respectively). When deployed, the multi-variable  
287 index process will allow users to specify a combination of indices. User-  
288 defined indices, also under development, will allow custom index calculations  
289 for single or multiple variables.

290

291 *Use case: Extremoscope.* We built an [interactive data visualization](#) to show  
292 the evolution of 10 extreme climate indices in the 13 regions of France under  
293 two climate scenarios (RCP 4.5 and 8.5) as part of the [Extremoscope](#) project.  
294 Flyingpigeon was used to calculate the climate indices using 7 climate pro-  
295 jection models for five time aggregates (four seasons and yearly, see: [4](#)).

296

297 The visualization displays the probability of extreme seasons or years oc-  
298 ccurring, defined as exceeding the 90th/95th percentile, or falling below the

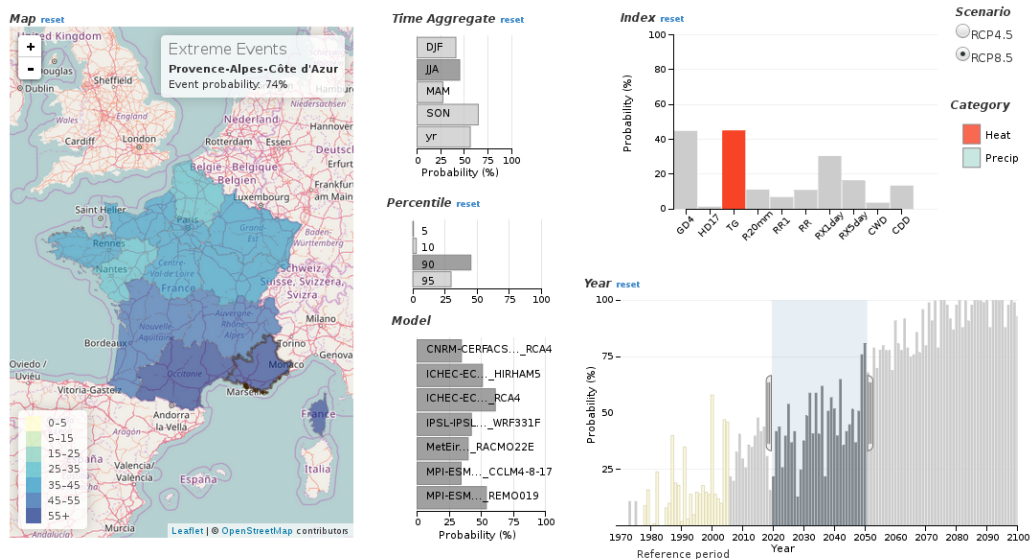


Figure 4: Screenshot of the Extremoscope visualization showing the probability averaged over all models of extreme (> 90th percentile) summer (JJA) temperatures over France from 2020-2050 under climate scenario RCP8.5.

299 5th/10th percentile, based on a reference period between 1972–2100. The  
 300 data is filterable by each dimension (region, time aggregate, index, scenario,  
 301 percentile, model and year) with different charts that are linked (crossfiltered)  
 302 using the [crossfiltering library](#) in [dc.js](#), the Dimensional Charting JavaScript  
 303 library, to allow the effects of each dimension to be observed and compared.  
 304

### 305 3.3. Climate impact related processes

306 One main target user group of the flyingpigeon is the climate impact  
 307 community. In this section, the currently embedded processes for general or  
 308 specific climate impact studies are presented. More processes like population  
 309 dynamics for *anopheles gambia*, the vector transmitting *Plasmodium falciparum*  
 310 to assess risks of malaria infection are under development and will be  
 311 implemented in upcoming versions.

#### 312 3.3.1. Segetal flora

313 'Segetal flora' is a term used for weeds growing in crop fields. The major-  
 314 ity of segetal flora species have very positive and important ecological effects

315 as a source of pollen or nectar for several insects or as nutrition for birds [16].  
316 Based on field monitoring, a relation between the number of different species  
317 occurring in crop fields and the annual mean temperature can be shown.  
318 The relations are investigated separately for seven segetal flora groups (e.g.  
319 Mediterranean or Nordic groups) and three land use types (conventional and  
320 ecological land use, and one–two year self-greened fallow land). The species  
321 number to mean annual temperature relations are expressed as regression  
322 functions [15, 14], which are the core of the segetal flora process. With tem-  
323 perature at 2m height (tas) as input data, and arguments to select segetal  
324 flora type and land use, the number of species is predicted. A polygon subset  
325 for a specified country can be selected with an optional argument.  
326 The process returns a tar file containing NetCDF files with the appropriate  
327 number of segetal flora species per grid point for each input dataset. The  
328 calculation is only relevant for areas with agricultural land use.

329

### 330 3.3.2. *Species distribution model*

331 Species distribution models (SDMs) are numerical tools to describe the  
332 relation between the distribution of a species and the environmental condi-  
333 tions that are thought to lead to this distribution [11, 24]. Models are  
334 used to gain ecological insights and predict distributions across landscapes,  
335 sometimes requiring extrapolation in space and time, e.g. for climate change  
336 impact modeling (modified from [8]).

337 We implemented a SDM process that relates the distribution of tree species to  
338 climate conditions to analyze the impacts of climate change [9]. The method  
339 focuses on tree species where the distribution is independent of non-climate  
340 factors like soil or specific site conditions. We hypothesize that the climate  
341 conditions which limit the tree species distribution (e.g. heat, precipitation,  
342 winter temperatures, drought [21]) can be described with climate indices (de-  
343 fined in section 3.2).

344 Fig. 5 shows the schematic workflow of the SDM process. Climate indices are  
345 calculated based on climate model variables (daily temperature and precipi-  
346 tation values). Tree species data can be fetched from the global biodiversity  
347 information facility (GBIF) database and fed to the SDM process. The  
348 geographical coordinates of the tree occurrence are then translated into a  
349 presence-absence (PA) matrix with the grid size of the input climate model  
350 data. Ocean and large lake areas are excluded in the PA matrix.

351 Statistical training [36, 35] is performed based on the calculated climate

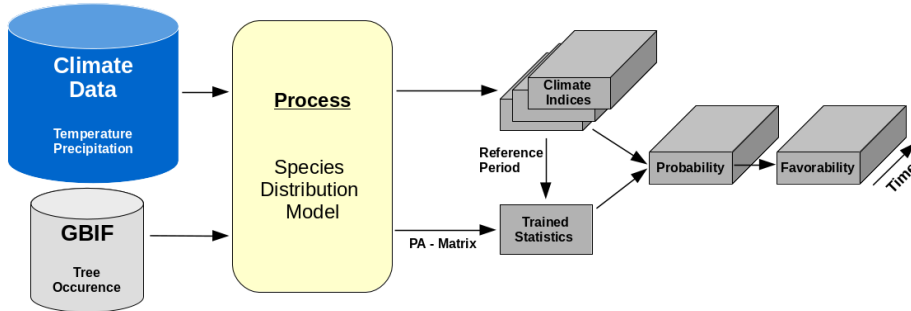


Figure 5: Schema of the Species Distribution Model process.

352 indices as a mean over a reference period and the PA matrix using general-  
 353 ized additive models (GAMs)[33, 37]. GAMs are flexible regression models  
 354 with non-parametric (smooth) additive components [34].

355 After the statistical training, the probability of occurrence is predicted over  
 356 the entire time series and transformed into a favorability [25] to compare dif-  
 357 ferent models. The SDM process can be used with different climate indices,  
 358 tree species, training performance values, and climate model datasets.

359

### 360 3.4. Extreme weather events related processes

361 Besides climate impact studies, the WPS flyingigeon also focuses on ex-  
 362 treme weather events investigations. This section contains examples of more  
 363 complex processes to calculate weather regimes (3.4.1) or analogues of circu-  
 364 lation (3.4.2).

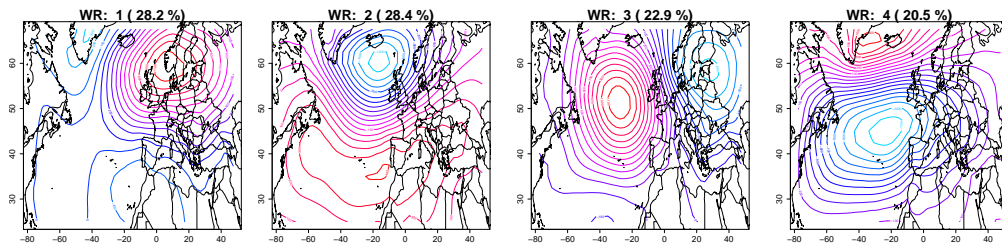


Figure 6: Weather regimes (WR) for winter months (DJF) based on NCEP sea surface pressure data (1970–2010).



365 *3.4.1. Weather regimes*

366 Weather regimes are recurring states of the atmosphere and provide a use-  
367 ful description of atmospheric variability [22, 6, 30]. In extreme event studies,  
368 for instance, it is possible to associate some specific seasonal weather regimes  
369 with the extremes of surface variables (precipitation or temperature) by an-  
370 alyzing which regime is prevailing when an extreme is encountered [40].

371 Following the methods of [22] and [39], the weather regime process computes  
372 a given number of regimes (the  $\kappa$  value, by default  $\kappa = 4$  [22]) for a region  
373 (by default the North Atlantic region [80°W - 50°E , 20°N - 70°N]) using  
374 values of sea level pressure (SLP) or geopotential height ( $Z_g(h)$ ) anomalies  
375 of a given season.

376 The weather regimes are computed with a k-mean classification algorithm  
377 [13] on the first 10 principal components (PCs) of the SLP or geopotential  
378 height ( $Z_g(h)$ ) fields. The PCs are computed from a Empirical Orthogonal  
379 Function decomposition [32] of the field. The data are weighted by the square  
380 root of the cosine of the latitude to account for grid cell surface variations.  
381 The obtained PCs are then classified onto  $\kappa$  clusters. The weather regimes  
382 can be computed on a reference data set (e.g. reanalyses data). Then, other  
383 datasets (e.g. climate model data) can be classified according to these refer-  
384 ence weather regimes.

385 Fig. 6 shows the output for a use case where weather regimes in winter  
386 months were calculated based on NCEP sea surface pressure data over a re-  
387 gion in the North Atlantic for  $\kappa = 4$ . The process also output an R workspace  
388 containing the statistical training values as well as a text file with the PCs  
389 and the NetCDF file with the normalized pressure values. These outputs can  
390 be used for manual postprocessing or for the weather regime projection pro-  
391 cess which calculates the percentage of each given weather regime in every  
392 year. Such a projection was done in this example for a CMIP5 dataset and  
393 Table 1 shows the calculated values.

394 *3.4.2. Analogues of circulation*

395 Analogues of circulation provide a versatile tool to investigate the rela-  
396 tion between climate variables (such as temperature or precipitation) and  
397 large-scale atmospheric circulation patterns (SLP or  $Z_g(h)$ ).

398 For an SLP/ $Z_g(h)$  pattern on a given day, the idea is to select days that have  
399 a calendar proximity (i.e. within a time window around the given date in  
400 all years except the year of the given day) and that minimize a distance be-  
401 tween the circulation patterns. This approach has been used to infer climate



Table 1: Percentual occurrence of weather regimes in winter months (DJF) trained on NCEP sea surface pressure data projected on a CMIP5 global dataset.

Year	WR 1	WR 2	WR 3	WR 4
...				
2084	35.16	28.57	30.77	5.49
2085	33.33	24.44	36.67	5.56
2086	21.11	28.89	40.00	10.00
2087	37.78	11.11	10.00	41.11
2088	18.68	19.78	37.36	24.17
2089	34.44	44.44	17.78	3.33
...				

402 reconstructions from SLP fields [26, 41], in weather forecast predictability as-  
 403 sessment [20], downscaling of climate variables [43, 29], detection/attribution  
 404 studies [31, 27, 4] and stochastic weather generators [38].

405

406 *Analogues detection process.* The analogues from this flyingpigeon process  
 407 are mainly used to study the conditional attribution of extreme events to dy-  
 408 namics, thus describing only the thermodynamic changes in weather events.  
 409 For a given continuous extreme event period (e.g. an extreme monthly tem-  
 410 perature), the process seeks in past archives flows similar to those in the  
 411 continuous period and reconstructs a monthly temperature from these past  
 412 flow analogue days. The resulting temperature is then compared to the cur-  
 413 rent one, providing an estimate of thermodynamic changes.  
 414 The process uses a circulation analogue simulation FORTRAN code ([CASTf90](#))  
 415 to compute analogues of circulation. The user can choose a time period for  
 416 which daily analogues are computed and an archive period, that is, the time  
 417 period to resample from. The user can also select a distance (Euclidean,  
 418 Mahalanobis, correlation, Teweles-Wobus S1 [28], how many analogues per  
 419 day to retain, a rectangular region (in longitude–latitude), whether to work  
 420 with anomalies or not and some more options like the output file format.  
 421 The result is a list of analogue dates for each simulation day and the corre-  
 422 sponding distance values.

423

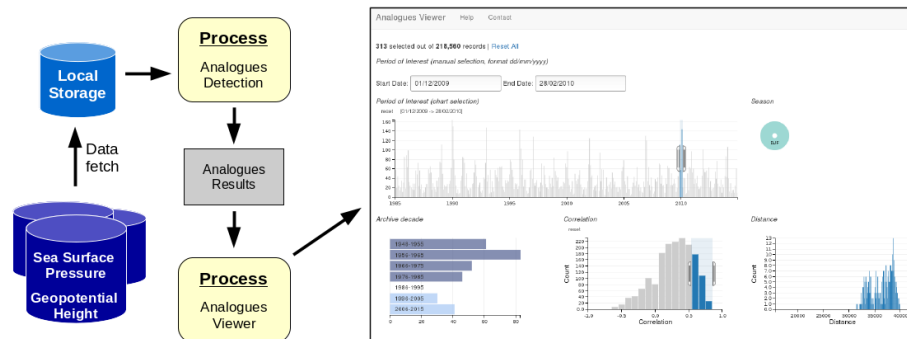


Figure 7: Schema of the analogues workflow.

424 *Analogues viewer process.* To explore the properties of the analogues detec-  
 425 tion process, we developed an analogues viewer process that produces an  
 426 interactive data visualization output as an HTML page hosted on the server-  
 427 side. This viewer uses [dc.js](#), the Dimensional Charting JavaScript library  
 428 and a [crossfiltering library](#) to produce interactive charts that can be filtered  
 429 based on data parameters. For each filter selection, all charts are simultane-  
 430 ously updated (crossfiltered).

431 Figure 7 shows a schematic of the analogues workflow with a screenshot of  
 432 the visualisation. For this example, the winter of 2010 was selected in the  
 433 bar plot, which filtered all charts to display counts of only those analogues  
 434 pertaining to this time period. The display was further filtered by selecting  
 435 analogues with correlation coefficients  $> 0.5$ . The decadal distribution of  
 436 analogues and changes across time can be immediately seen by sliding the  
 437 time window.

438 The analogues viewer demonstrates two powerful advantages of using WPS  
 439 for data analysis: 1) it is straight-forward to 'chain' processes together so  
 440 that the output of one process can be used as the input of another, allowing  
 441 data analysis to be re-imagined as a workflow, and 2) processes can be com-  
 442 pletely custom-designed. Here, we integrate data visualization capability –  
 443 normally a step performed in an environment completely separate from the  
 444 analysis and requiring a specialized skill that is often not within the scope  
 445 of the scientific researcher – directly into flyingpigeon, accessible simply by  
 446 launching the output HTML link.

447

#### 448 **4. Conclusion**

449 Flyingpigeon, as a compartment of birdhouse, is an advanced toolbox  
450 that can be used to process the increasingly large amount of climate model  
451 data in a standardized and secure way. This current version is an initial se-  
452 lection of processes commonly used for climate impact and extreme weather  
453 event analysis. Flyingpigeon, connected to other OWSs, has access to various  
454 convenient features such as ESGF data search, visualisation via the WMS,  
455 and catalog services. It allows collaborative data and methods sharing and  
456 access to several data archives. Process outputs can be recycled in other  
457 processes in a workflow. All these advantages are possible regardless of the  
458 script language of the methods.

459 The transparency of the source code as an open source project enables quick  
460 and easy exchange of developer knowledge, good quality control and frequent  
461 updates of the analysis methods and performance improvements.

462 Flyingpigeon reduces the difficulty of data processing and is a solution to  
463 facilitate the daily work for the climate community.

464

#### 465 **5. Outlook**

466 The increasing number of developers and users ensures further improv-  
467 ment of features, documentation, guidelines and tutorials. Currently, high-  
468 performance computing providers located close to the data archives are being  
469 established in the climate community to keep up with the demand for WPS.

470

#### 471 **Acknowledgments**

472 NCEP Reanalysis data was obtained from NOAA/OAR/ESRL PSD,  
473 Boulder, Colorado, USA, from their website [http://www.esrl.noaa.gov/](http://www.esrl.noaa.gov/psd/)  
474 [psd/](http://www.esrl.noaa.gov/psd/).

475 20th Century Reanalysis V2 data was obtained from NOAA/ OAR/ESRL  
476 PSD, Boulder, Colorado, USA, from their website [http://www.esrl.noaa.](http://www.esrl.noaa.gov/psd/)  
477 [gov/psd/](http://www.esrl.noaa.gov/psd/).

478 CAC, SR and PY are supported by the ERC Advanced Grant No. 338965-  
479 A2C2.

480 **References**

- 481 [1] G. Brandl and Sphinx team. Sphinx. python documentation generator,  
482 2016.
- 483 [2] J. Brauner. *Formalizations for geoperators-geoprocessing in Spatial*  
484 *Data Infrastructures*. Dissertation, Technische Universität Dresden,  
485 Fakultät Umweltwissenschaften, 10 2015.
- 486 [3] A. M. Castronova, J. L. Goodall, and M. M. Elag. Models as web services  
487 using the open geospatial consortium (ogc) web processing service (wps)  
488 standard. *Environmental Modelling & Software*, 41(0):72–83, 2013.
- 489 [4] J. Cattiaux, R. Vautard, C. Cassou, P. Yiou, V. Masson-Delmotte, and  
490 F. Codron. Winter 2010 in europe: A cold extreme in a warming climate.  
491 *Geophysical Research Letters*, 37(20):L20704, 2010.
- 492 [5] G. P. Compo et al. The twentieth century reanalysis project. *Q. J. Roy.*  
493 *Meteor. Soc.*, 137(654):1–28, 2011.
- 494 [6] S. Corti, F. Molteni, and T. N. Palmer. Signature of recent climate  
495 change in frequencies of natural atmospheric circulation regimes. *Nature*,  
496 398(6730):799–802, 1999.
- 497 [7] C. Déandreis, C. Pagé, P. Braconnot, L. Bärring, E. Bucchignani, W. S.  
498 de Cerff, R. Hutjes, S. Joussaume, C. Mares, S. Planton, and M. Plieger.  
499 Towards a dedicated impact portal to bridge the gap between the impact  
500 and climate communities : Lessons from use cases. *Climatic Change*,  
501 125(3):333–347, 2014.
- 502 [8] J. Elith and J. R. Leathwick. Species distribution models: Ecological  
503 explanation and prediction across space and time. *Annual Review of*  
504 *Ecology, Evolution, and Systematics*, 40(1):677–697, 2009.
- 505 [9] W. Falk and N. Hempelmann. Species favourability shift in Europe due  
506 to climate change: A case study for *fagus sylvatical.* and *picea abies(l.)*  
507 *karst.* based on an ensemble of climate models. *Journal of Climatology*,  
508 2013:1–18, 2013.
- 509 [10] G. Fenoy, N. Bozon, and V. Raghavan. Zoo-project: the open wps  
510 platform. *Applied Geomatics*, 5(1):19–24, 2013.

- 511 [11] J. Franklin. *Mapping species distributions: spatial inference and predic-*  
512 *tion*. Cambridge University Press, 2010.
- 513 [12] M. Harrower and M. Bloch. Mapshaper.org: A map generalization web  
514 service. *IEEE Comput. Graph. Appl.*, 26(4):22–27, July 2006.
- 515 [13] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering  
516 algorithm. *Journal of the Royal Statistical Society. Series C (Applied*  
517 *Statistics)*, 28(1):100–108, 1979.
- 518 [14] N. Hempelmann and C. Ehbrecht. Web processing services for climate  
519 data - with examples for impact modelers. In *EGI Community Forum*  
520 *2014*, 2014.
- 521 [15] J. Hoffmann, N. Hempelmann, M. Glemnitz, L. Radics, G. Czimber, and  
522 U. Wittchen. Einfluss von temperatur und nutzung auf die floristische  
523 artenvielfalt in getreideanbaugebieten europas. 436:70-76. *Julius-Kühn-*  
524 *Archiv*, 2012.
- 525 [16] J. Hoffmann, U. Wittchen, N. Hempelmann, M. Glemnitz, and  
526 L. Radics. Wildpflanzen auf dem acker – möglichkeiten und grenzen für  
527 den artenschutz im Ökolandbau. *Forschungs Report Spezial Ökologischer*  
528 *Landbau*, page 20.21, 2013.
- 529 [17] C. Jung, M. Gasthuber, A. Giesler, M. Hardt, J. Meyer, F. Rigoll,  
530 K. Schwarz, R. Stotzka, and A. Streit. Optimization of data life cy-  
531 cles. *Journal of Physics: Conference Series*, 513(3):032047, 2014.
- 532 [18] E. Kalnay et al. The NCEP/NCAR 40-year reanalysis project. *Bulletin*  
533 *of the American Meteorological Society*, 77(3):437–471, 1996.
- 534 [19] S. Kindermann, F. Schintke, and B. Fritsch. A collaborative data man-  
535 agement infrastructure for climate data analysis. *Geophysical Research*  
536 *Abstracts*, 14(EGU201):10569, April 2012.
- 537 [20] E. N. Lorenz. Atmospheric predictability as revealed by naturally oc-  
538 ccurring analogues. *Journal of the Atmospheric Sciences*, 26(4):636–646,  
539 1969.

- 540 [21] K. H. Mellert, V. Deffner, H. Küchenhoff, and C. Kölling. Modeling  
541 sensitivity to climate change and estimating the uncertainty of its im-  
542 pact: A probabilistic concept for risk assessment in forestry. *Ecological*  
543 *Modelling*, 316:211–216, Nov 2015.
- 544 [22] P.-A. Michelangeli, R. Vautard, and B. Legras. Weather regimes: Re-  
545 currence and quasi stationarity. *J. Atmos. Sci.*, 52(8):1237–1256, 1995.
- 546 [23] M. Mueller and B. Pross. OGC® WPS 2.0 interface standard corrigendum 1, 03 2015.  
547
- 548 [24] A. T. Peterson, J. Soberón, R. G. Pearson, R. P. Anderson, E. Martínez-  
549 Meyer, M. Nakamura, and M. B. Araújo. *Ecological Niches and Geo-*  
550 *graphic Distributions (MPB-49) (Monographs in Population Biology)*.  
551 Princeton University Press, 2011.
- 552 [25] R. Real, A. M. Barbosa, and J. M. Vargas. Obtaining environmen-  
553 tal favourability functions from logistic regression. *Environ Ecol Stat*,  
554 13(2):237–245, Jun 2006.
- 555 [26] F. Schenk and E. Zorita. Reconstruction of high resolution atmospheric  
556 fields for northern europe using analog-upscaling. *Climate of the Past*,  
557 8(5):1681–1703, 2012.
- 558 [27] P. A. Stott, N. Christidis, F. E. L. Otto, Y. Sun, J.-P. Vanderlinden,  
559 G. J. van Oldenborgh, R. Vautard, H. von Storch, P. Walton, P. Yiou,  
560 and F. W. Zwiers. Attribution of extreme weather and climate-related  
561 events. *WIREs Climate Change*, 7:23–41, 2016.
- 562 [28] S. Teweles and H. B. Wobus. Verification of prognostic charts. *Bulletin*  
563 *of the American Meteorological Society*, 35(10):455–463, 1954.
- 564 [29] P. Vaittinada Ayar, M. Vrac, S. Bastin, J. Carreau, M. Déqué, and  
565 C. Gallardo. Intercomparison of statistical and dynamical downscaling  
566 models under the euro- and med-cordex initiative framework: present  
567 climate evaluations. *Climate Dynamics*, 46(3):1301–1329, 2016.
- 568 [30] R. Vautard. Multiple weather regimes over the north atlantic: Analysis  
569 of precursors and successors. *Monthly weather review*, 118(10):2056–  
570 2081, 1990.

- 571 [31] R. Vautard and P. Yiou. Control of recent european surface climate  
572 change by atmospheric flow. *Geophysical Research Letters*, 36:L22702,  
573 2009.
- 574 [32] H. Von Storch and F. W. Zwiers. *Statistical analysis in climate research*.  
575 Cambridge university press, 2001.
- 576 [33] S. Wood. *Generalized Additive Models: An Introduction With R*, vol-  
577 ume 66. CRC Press, 2006.
- 578 [34] S. N. Wood. Modelling and smoothing parameter estimation with mul-  
579 tiple quadratic penalties. *Journal of the Royal Statistical Society (B)*,  
580 62(2):413–428, 2000.
- 581 [35] S. N. Wood. Stable and efficient multiple smoothing parameter estima-  
582 tion for generalized additive models. *Journal of the American Statistical*  
583 *Association*, 99(467):673–686, 2004.
- 584 [36] S. N. Wood. Fast stable restricted maximum likelihood and marginal  
585 likelihood estimation of semiparametric generalized linear models. *Jour-*  
586 *nal of the Royal Statistical Society (B)*, 73(1):3–36, 2011.
- 587 [37] S. N. Wood and N. H. Augustin. Gams with integrated model selec-  
588 tion using penalized regression splines and applications to environmental  
589 modelling. *Ecological Modelling*, 157(2-3):157–177, Nov 2002.
- 590 [38] P. Yiou. AnaWEGE: a weather generator based on analogues of at-  
591 mospheric circulation. *Geoscientific Model Development*, 7(2):531–543,  
592 2014.
- 593 [39] P. Yiou et al. Weather regime dependence of extreme value statistics  
594 for summer temperature and precipitation. *Nonlinear Proc. Geoph.*,  
595 15(3):365–378, 2008.
- 596 [40] P. Yiou and M. Nogaj. Extreme climatic events and weather regimes  
597 over the north atlantic: When and where? *Geophysical Research Letters*,  
598 31(7), 2004.
- 599 [41] P. Yiou, T. Salameh, P. Drobinski, L. Menut, R. Vautard, and M. Vrac.  
600 Ensemble reconstruction of the atmospheric column from surface pres-  
601 sure using analogues. *Climate Dynamics*, 41:1333–1344, 2013.

- 602 [42] S. Zhou and C. B. Jones. *Developments in Spatial Data Handling:*  
603 *11th International Symposium on Spatial Data Handling*, chapter Shape-  
604 Aware Line Generalisation With Weighted Effective Area, pages 369–  
605 380. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.
- 606 [43] E. Zorita and H. von Storch. The analog method as a simple statistical  
607 downscaling technique: comparison with more complicated methods.  
608 *Journal of Climate*, 12(8):2474–2489, 1999.