

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.Doi Number

# Exploring Stigmergic Collaboration and Task Modularity through an Expert Crowdsourcing Annotation System: The Case of Storm Phenomena in the Euro-Atlantic Region

DENNIS PAULINO<sup>1</sup>, ANTÓNIO CORREIA<sup>1,2</sup>, MARCELA MAYUMI MAURICIO YAGUI<sup>3</sup>,  
JOÃO BARROSO<sup>1</sup>, MARGARIDA L. R. LIBÉRATO<sup>4</sup>, ADRIANA S. VIVACQUA<sup>3</sup>,  
ANDREA GROVER<sup>2</sup>, JEFFREY P. BIGHAM<sup>5</sup>, AND HUGO PAREDES<sup>1</sup>

<sup>1</sup> INESC TEC and University of Trás-os-Montes e Alto Douro, Vila Real, Portugal

<sup>2</sup> University of Nebraska at Omaha, Omaha, NE, United States

<sup>3</sup> Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

<sup>4</sup> IDL and University of Trás-os-Montes e Alto Douro, Vila Real, Portugal

<sup>5</sup> Carnegie Mellon University, Pittsburgh, PA, United States

Corresponding author: Dennis Paulino (dpaulino@utad.pt)

This work is financed by the Portuguese Foundation for Science and Technology (Fundação para a Ciência e a Tecnologia – FCT) with the research grant SFRH/BD/148991/2019. The authors also acknowledge support from the European Social Fund under the scope of North Portugal Regional Operational Programme.

**ABSTRACT** Extreme weather events, such as windstorms, hurricanes, and heat waves, exert a significant impact on global natural catastrophes and pose substantial challenges for weather forecasting systems. To enhance the accuracy and preparedness for extreme weather events, this study explores the potential of using expert crowdsourcing in storm forecasting research through the application of stigmergic collaboration. We present the development and implementation of an expert Crowdsourcing for Semantic Annotation of Atmospheric Phenomena (eCSAAP) system, designed to leverage the collective knowledge and experience of meteorological experts. Through a participatory co-creation process, we iteratively developed a web-based annotation tool capable of capturing multi-faceted insights from weather data and generating visualizations for expert crowdsourcing campaigns. In this context, this article investigates the intrinsic coordination among experts engaged in crowdsourcing tasks focused on the semantic annotation of extreme weather events. The study brings insights about the behavior of expert crowds by considering the cognitive biases and highlighting the impact of existing annotations on the quality of data gathered from the crowd and the collective knowledge generated. The insights regarding the crowdsourcing dynamics, particularly stigmergy, offer a promising starting point for utilizing stigmergic collaboration as an effective coordination mechanism for weather experts in crowdsourcing platforms but also in other domains requiring expertise-driven collective intelligence.

**INDEX TERMS** Atmospheric phenomena, cognitive biases, crisis informatics, expert crowdsourcing, extreme meteorological events, semantic annotation, stigmergy, storms, task modularity, weather maps.

## I. INTRODUCTION

Withstanding unpredictable weather events is one of the biggest perennial challenges facing community ecology [1]. The harsh impacts of severe meteorological events such as extreme windy conditions and precipitation extremes are increasingly being felt and pose serious threats to ecosystem functioning [2]. Such extreme weather phenomena resulting from climate change can inflict

profound socio-economic impacts, increasing economic losses, agriculture production shortfalls, and human mortality [3, 4]. These factors, in tandem with the magnitude, frequency, and duration of these extreme events, have prompted a call for collective action to enable early detection of abnormal weather conditions in a way that can be useful to prevent potential hazards across the globe [5]. Multi-hazard early warning systems expose

significant weaknesses related to the uncertain, multifaceted and unexpected nature of severe hazards caused by weather extremes [6]. Since these events are exceptions to well-known weather patterns, they are not forecasted appropriately. As climate change accelerates and catastrophic weather events become more frequent, intense, and unpredictable [7], innovative ways must be found to improve weather pattern detection for extreme weather events around the world.

Despite the utter importance of crisis informatics to the human-computer interaction (HCI) research community [8-10], many aspects of crisis preparedness, extreme weather warning response, and post-disaster recovery through collaborative and social computing have not yet been investigated intensively. As outlined above, turning weather observations (e.g., annotations) provided by human contributors into reusable data that can be further interpreted by weather forecast systems constitutes a long-standing problem [11]. There is much accumulated evidence that both expert and non-expert crowds can be reliable sources of observational data in the form of semantic annotations of image fragments [12], even though the accuracy and inter-annotator agreement varies depending on the task structure [13].

That brings us to the issue of task modularity [14] to reduce the coordination burden. Typically, conventional microtask crowdsourcing platforms leverage non-expert crowd workers to perform such tasks based on a random task distribution strategy [15]. However, analyzing extreme weather events is not a simple activity that can be easily decoupled from the expertise of specialists like climatologists and atmospheric researchers. With the increasing global concern for the complex environmental issues associated with the effects of extreme weather and climate events, the field of Big Data has an important role to play in supporting the development of technologies for extreme weather detection and prediction, based on its strong tradition of investigating situated work practices from a multidisciplinary perspective [16, 17]. This is accentuated by the recent evidence that making predictions about the future occurrence of extreme event phenomena based on historical data (i.e., the frequency of extreme events occurred in the past) fall short in supporting desired levels of effectiveness [18].

Niforatos and colleagues [19] suggested that a human-in-the-loop approach can be particularly useful to estimate weather conditions at multiple points while training machine learning algorithms into a dynamic process that goes beyond the common conception of human oversight that has been widely used in the field of artificial intelligence [20]. Evidence of such a process in a slightly different domain has recently been reported by Kim and Pardo [21], who claimed that algorithmic sociotechnical systems leveraged by human-in-the-loop interactions can progressively improve the speed and quality of the labeling

process while reducing manual effort, with promising results in the context of sound event annotation. Therefore, we believe that the involvement of multiple researchers with expertise in climate-related application domains can contribute to a richer and more efficient analysis of adversarial weather phenomena when working collaboratively. Crowdsourcing is often used as an engagement solution, allowing an efficient usage of human computation [22]. Muller and co-authors [23], referring to Wechsler [24], briefly explained the applicability of this approach in climate and atmospheric sciences as follows: “Crowdsourcing as a research field has great potential to bridge the gap between the social scientists, computer scientists and physical and environmental scientists, thereby encouraging interdisciplinary working and enhancing knowledge exchange and scientific discovery”.

In this paper, we examine the application of crowdsourcing in climate and climate change research, through expert crowd annotation of atmospheric phenomena, following similar approaches such domains as oceanography [25]. In particular, our work asks whether expert crowdsourcing is effective for detecting extreme weather events, in tandem with the aim of exploring the potential effects of task modularity [14]. Furthermore, we also report data involving stigmergy, a form of implicit coordination in which influence is exerted by indirect interactions via a self-organizing environment based on digital traces left by participants without apparent communication or separate articulation work [26, 27]. The idea here is that since social coordination is particularly challenging to manage in uncertain scenarios [28], expert crowds can collaboratively tackle complex problems such as extreme weather event detection and climate forecasting by adjusting behaviors based on past contributions provided by other members. As Bolici et al. [26] put it, a stigmergic approach considers the shared context itself, including material artifacts that are part of the environment, as a coordination mechanism. This allows a sharing of lessons learned and a common design ground, which is in line with a central question for the computer-supported cooperative work (CSCW) research community concerning the best ways of promptly and accurately obtaining crowdsourced annotations [29] and how they are collaboratively achieved in everyday practice. Through a crowdsourcing process, with tools developed for mapping the collaborative process in the crowdsourcing framework, climate experts were able to participate in online tasks following a collaborative problem-solving approach. The potential advantages of leveraging human factors in weather forecasting systems include improving their accuracy and reliability by incorporating knowledge and experience from experts. This knowledge can be then turned into a set of actionable insights that are left in the environment to support shared awareness while providing non-professionals a better understanding of weather phenomena as informed by experts. In line with this, our

study aims to address research gaps in this area and contribute to a better characterization of how expert crowds can be effectively coordinated to enhance weather forecasting systems.

The remainder of the paper is structured as follows. In section 2 we recapitulate the related literature on stigmergic collaboration, citizen science, and expert crowdsourcing, with an emphasis on the advantages of semantic annotation afforded by web-enabled mass collaboration. In section 3 we describe the characteristics of the crowdsourcing platform created for supporting the annotation of extreme weather events. In this section, we also formally model the system components, alongside a brief description of the user interface and related tools used in our study. Section 4 describes the data collection and processing procedures, outlining in detail the methods used to analyze the qualitative data collected. In section 5, we present the experimental results obtained from real-world data. These findings compare the data obtained from the annotation experiments and investigate the relative importance of stigmergy and task modularity on the general outputs provided by experts, with focus on the agreement among participants as a measure of collective intelligence. In section 6, we discuss these findings in light of previous studies with a view to the current uses of the platform. This section also discusses some limitations of our research and brings out lessons that have been learned during this study in the form of implications for design. We close with a review of our findings and a proposal for further research.

## II. BACKGROUND AND RELATED WORK

### A. The Metaphor of Stigmergic Collaboration

Stigmergy in crowd work relies on tracking changes on the work activity under execution using digital traces as guidance mechanisms for allowing other members to contribute based on past behavior [30]. In many ways, it is analogous to the trails left by insects as a self-organizing, co-evolving approach in which the previously performed work plays a critical role by acting as an indirect or even spontaneous coordination mechanism [24, 31]. This is in line with Bolici and associates [26], who claimed that implicit coordination is “reached without discursive communication, shared plans or even previous commitment among the actors”. In this kind of scenario, both the shared artifact and the environment act as a source of stimuli to the ensuing interactions in situated actions that are mainly characterized by ad hocness [32]. In [33], for example, a stigmergic system approach was introduced to support crowd-powered decision-making and reasoning tasks with emphasis on aiding the work of intelligence analysts without the use of explicit forms of coordination.

Another well-known example of this stigmergic approach is Wikipedia [30], where contributors seem to be working

separately but in fact they are triggered in part by the signs left by other members in the shared environment without control of each other [34]. This results in a complex network constituted by a long-term shared external working memory in which the users are attracted to participate and build upon an initial concept or idea as a form of collective intelligence [26, 35-37]. Crowston and Rezgui [30] found a positive influence of non-explicit, stigmergic coordination on the quality of a set of articles produced in Wikipedia. In such settings, task interdependence and worker roles do not affect the success of collaboration [21].

In addition, cognitive biases may also affect stigmergy-based approaches and thus cause individuals to make mistakes without being aware of them. From a conceptual level, cognitive biases can be understood as the systematic production of distorted representations that are detached from reality by an individual's own cognition [38]. As the technologies enabling crowdsourcing options have evolved and matured over the years, a body of evidence has accrued indicating that the quality of tasks can significantly decrease when cognitive biases are not taken into account in the task design process [39]. Among many types of cognitive biases that can influence crowd behavior, the Dunning-Kruger effect [40] occurs when less skilled individuals who perform at levels below expectations on a particular task tend to overestimate their performance. For instance, a study based on input aggregation methods for crowdsourcing systems [41] demonstrated that the general performance of the aggregation methods achieved better results when the Dunning-Kruger effect was taken into consideration. In related work, Gadiraju et al. [42] examined a set of crowd workers' self-assessments and found that less diligent workers tended to be more affected by this type of bias. As a result, we believe that stigmergic processes could benefit from assessment for individuals' cognitive biases to improve the collaboration outcomes that may be achieved in this form of implicit interaction. Further, the Bandwagon effect [43] can also be particularly relevant in a stigmergy-based scenario.

### B. Citizen Science and Expert Crowd Work for Weather Event Detection

Crowd science, networked science, massively collaborative science, or simply citizen science are types of crowdsourcing that encourage public participation in doing scientific research, for example, by contributing observations, historical records, and collection of samples and scientific evidence [44]. Citizen science encourages concerned people to actively participate in a real science project alongside professional researchers [45]. Citizen science projects usually enable different levels of crowd participation based on the project goals [46]. In the literature, many studies applying a citizen science strategy have been published in the natural sciences [47]. For instance, current initiatives promote

citizens' engagement in addressing climate change through the usage of games and social networks [48].

Regarding task complexity, Franzoni and Sauermann [44] claims that interconnection among tasks increases the degree of complexity and Staffebach, Sempolinski [49] argues that task complexity is due to the technical knowledge that the task requires. In [50], the authors relate task complexity to the degree of collaboration among participants. In [51] created a method to identify profiles and relate them to the most appropriate tasks, while Cullina and co-authors [52] proposed a sub-process to select the best crowd for crowdsourcing projects. In [53], the authors studied the reliability and cost of approaches such as 'majority decision' and 'control group' for validating crowd contributions. We also note the work of De Beer et al. [54] on the ethical issues related to the intellectual property of data generated by crowdsourcing systems from the perspective of the organizations carrying out crowdsourcing activities as requesters.

#### 1) EXPERT CROWDS VS. AMATEURS (NON-PROFESSIONALS)

Some studies have considered the characteristics of contributors and their interactions for assessing the quality of an artifact [55]. That is, the interactions among crowd members are inherently contextual and situated [56], which influences the general quality of the crowd work product. A mismatch in the quality of volunteer contributions can occur when the views of professional experts (scientists) differ from non-experts and vice-versa [57, 58]. Complex tasks must be assigned to an expert taking into consideration their particular skills and abilities [27], despite multiple barriers to high-quality expert-derived annotations, particularly as reliance on expert contributors is time-consuming and expensive [59].

On the other hand, non-experts may experience more difficulties in performing advanced complex tasks, potentially due to lack of field-specific expertise. Increasingly, research comparing the performance of volunteers (also known as citizen scientists) against the work done by scientific experts [60] has pushed the boundaries in this direction. In particular, O'Leary and colleagues [61] tested the validity of using non-experts for collecting phenotypic data. Current works have similarly evaluated the performance of domain experts, crowd workers, and automated techniques (e.g., active learning) for sleep spindle detection [62] and finite-pool data categorization of datasets in biomedical systematic reviews [63]. Moreover, Law and co-authors [64] also identified some barriers associated with the hesitancy of experts when using data collected from citizen scientists. In this regard, [65] emphasized recursive problems such as the absence of a shared disciplinary background and lack of trust when considering the perceptions of crowd bias and data quality.

#### 2) CROWDSOURCING ATMOSPHERIC DATA & DATA ANALYSIS.

A few studies have attempted to use stigmergic principles in the study of crowdsourcing for annotating atmospheric phenomena. In crowdsourcing for identifying the formation of strong tropical cyclones, Barbosa and Gans [66] examined the use of Cyclone Center<sup>1</sup>, a Zooniverse project that recruits volunteers for analyzing global hurricane records using satellite imagery. In a similar vein, Hennon et al. [67] also explored the potential of using this crowdsourcing system for identifying patterns in tropical cyclone analysis and forecasting. In [68], the authors applied crowdsourcing for analyzing air temperature data (e.g., cold and heat anomalies) retrieved from Internet-connected weather stations. In an entirely different approach, citizen science was utilized to involve the broader public in providing volunteer computing resources for supporting climate studies [69]. In another citizen science project<sup>2</sup>, volunteers are invited to transcribe meteorological readings on past environmental conditions [70]. Crowdsourcing transcripts of old atmospheric records was also explored within a traceable citizen science-based environment and climate database [71].

Data from a variety of additional citizen science projects, are increasingly incorporated into standard data sets used in atmospheric sciences, with volunteers gathering observational data for certain meteorological parameters. For example, the Weather Observation Website<sup>3</sup> was initiated as a citizen science project intended to crowdsource weather data from private observers with the aim of building up a record of weather observations [72]. Researchers have also examined the amateur weather observation practices and experiences from a sociotechnical viewpoint [72]. For a detailed overview of previous work on the challenges and possibilities of using crowdsourcing for meteorological studies, we refer the reader to [23].

#### C. Semantic Annotation via Crowdsourcing

Research has shown that semantic annotations can facilitate the process of attaching additional information to various concepts so that algorithms can automatically process and interpret them effectively, using structured metadata available on the web to associate terms or expressions in a document with an instance of an ontology [73, 74]. In a recent study by Hughes et al. [75], a crowd-powered tool called Quanti.us leveraged the wisdom of untrained crowd workers for semantically annotating scientific images. Assessments confirmed similar performance for broader public and expert-derived annotations in terms of precision and recall, with a slightly higher accuracy rate for experts on more complex annotation tasks. For semantic annotation tasks, some works (e.g., [76-78]) were developed with the objective of annotating metadata on maps via crowdsourcing techniques. Kaufmann et al. [77] proposed the creation of an

<sup>1</sup> <https://www.cyclonecenter.org/>

<sup>2</sup> <https://www.oldweather.org/>

<sup>3</sup> <https://www.metoffice.gov.uk/>



application, called Climate Twins, that simulates the future climate of some regions of Europe, based on historical series of temperature and precipitation.

Semantic annotations can also be combined with other semantic web techniques such as Linked Open Data (LOD) [79] to facilitate the publication of structured data on the Web and the creation of connections between heterogeneous data sources. For example, the work of Santos and Furtado [78] relies on the creation of annotated maps based on content generated by virtual crowds, where users can make markings on maps and describe events. Work by Gonzalez and colleagues [76] has also proved the applicability of crowdsourcing to produce LOD datasets using map annotations[80]. The authors tested their proof-of-concept approach in the geospatial domain. In such scenarios, the descriptions can be obtained from the LOD cloud [81] where the curated information on the topics is made available. In a different domain, Callaghan et al. [82] explored the use of crowdsourcing for annotating heart sounds with excellent classification performance when hybrid machine-crowd classifiers are used.

### III. Technical Setup

In this section, we draw on the recent success of crowd-powered systems to support annotation tasks (e.g., [29, 83, 84]) to introduce the expert Crowdsourcing for Semantic Annotation of Atmospheric Phenomena (eCSAAP) tool, a technical solution to the problem of finding patterns in weather data for forecasting. This section discusses elicitation of requirements and associated system features.

The development and design phases followed a participatory design approach [85] to understand requirements of atmospheric specialists, meteorologists, and climate scientists when collecting data about extreme weather phenomena. During this process, we talked with experts in climatology, working in the domain of extreme events (i.e., extratropical cyclones), about the types of weather data they usually work with and the way they interact with these data, including possible barriers and limitations with these arrangements. We adopted an iterative-incremental development strategy based on a co-creation process, where requirements evolved through the interaction and direct feedback over a 9-month development period (March – November 2019). This process informed the design of a prototype “from the ground up” [86] as opposed to using an existing tool. The goal of this process was to easily modify each feature every time a change was required in the sequence of the feedback received by expert participants.

To support the needs of climate specialists, we required a web-based tool that would adequately provide the output data in the form of marks (annotations) at specified locations as rich content-based descriptions of the observed weather phenomena over climatological time spans. This imposed an additional requirement that the platform should be developed to support overlapping annotations, as also shown in earlier

studies (e.g., [29]). Another system requirement was to provide a mechanism to enable the visualization of multifaceted insights from automatically generated weather maps, which is highly desirable so that the meteorological parameters can be defined for each task. As a result, the work units that are crowdsourced must be modular in a way that simplifies keeping them compliant with the flow of information, interactions, and collaboration while preserving stakeholder privacy and security. Based on the requirements elicited during the participatory design activity, we designed a framework to annotate weather events. At first glance, the eCSAAP-Service Oriented Framework for Weather Annotation (eCSAAP-SOF4WA) encloses four functional blocks (Figure 1): Weather Center API, eCSAAP System, Extreme Weather Events Annotator, and Crowdsourcing Platform. The functional blocks are organized in order to provide an end-to-end approach to classify weather phenomena, gathering weather data from international databases, generating experts-oriented visualizations and attaching semantic fields to the collected data, and creating tailored crowdsourcing campaigns for weather annotation.

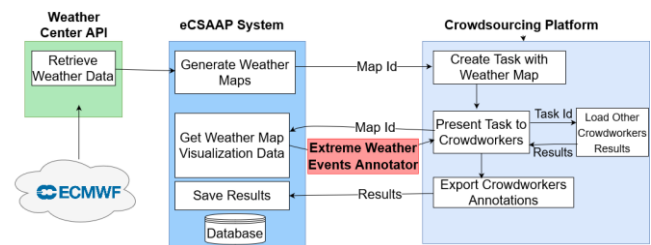


FIGURE 1. Overview of eCSAAP-SOF4WA.

#### A. Weather Center API

The Weather Center API provides weather data to the eCSAAP system by abstracting the data sources. Several weather data sources currently provide weather data gathered directly from weather sensors or through reanalysis. The current implementation of the eCSAAP-SOF4WA gathers reanalysis data from the European Centre for Medium-Range Weather Forecasts (ECMWF)<sup>4</sup>. Such data are publicly available at a variety of scales, derived from a diverse set of meteorological observations that enable us to support weather forecasting activities based on the current state of the atmosphere.

#### B. eCSAAP System

From a sociotechnical perspective, the eCSAAP system can be used by experts and non-professionals (citizen scientists, weather enthusiasts, students) worldwide to generate visualizations from atmospheric data [omitted for blind review]. Interoperability was taken into account in the deployment of the eCSAAP system through development of two components: a backend (management of requests to the

<sup>4</sup> <https://www.ecmwf.int/>

database and production of meteorological views) and a frontend that consists of an interface for users to interact with the backend. The backend was built using Django<sup>5</sup>, a Python-based framework for developing a Restful API, which works with map libraries such as Matplotlib<sup>6</sup> and makes the generation of visualization maps easier. The main features of the backend are briefly described below:

- *Collection of meteorological data from databases.* A set of parameters such as the space-time interval and meteorological metrics are sent to an API. The data returned is part of reanalysis, which allows a consistent global view for metrics such as wind and mean sea level pressure [87]. Weather data is collected using the Climate Data Store API<sup>7</sup>, implemented by the ECMWF.
- *Generation of weather maps and visualization data.* The collected weather data are represented on two-dimensional maps, which are projected on a world map. When creating each map, visualization data is stored (i.e., geometric objects coordinates representing a weather phenomenon). This visualization data is available as a resource through the backend API, and is used by the Extreme Weather Events Annotator described in Section 3.3 to create interactive maps.
- *Crowdsourcing task management.* This feature allows users to create crowdsourcing campaigns by making HTTP requests to a crowdsourcing platform API. When a task is created, it is associated with multiple maps, allowing the monitoring of the evolution of a meteorological phenomenon such as a storm. After the campaign is finished in the crowdsourcing platform, then the results can be stored for further analysis.

Weather experts can use the backend through a user-friendly frontend. The frontend was developed with React<sup>8</sup>, which facilitates the consumption of APIs. The weather experts can do every task available through the backend, from generating weather maps to creating crowdsourcing campaigns for weather annotation.

### C. Extreme Weather Events Annotator

A web-based annotator allows users to create semantic annotations of extreme weather events, as shown in Figure 2. The annotator was developed in HTML5 and Javascript to facilitate the integration of the task interface with the crowdsourcing platform. The task interface consists mainly of an interactive Leaflet<sup>9</sup> map where it is possible to draw various geometric objects. In this map, there are overlays that represent the visualization of meteorological metrics,

which can be projected in the form of colored polygons or polylines (Label 1 in Figure 2).

The geometric object style (e.g., color or line style) can be customized to show more information with a popup presenting the value of the weather metric (Label 2 in Figure 2). Annotations can be created in two ways: by clicking on a geometric object to choose the rectangular area (Label 3 in Figure 2) or by manually drawing a rectangle in the desired area (Label 4 in Figure 2). Each annotation can be edited (Label 5 in Figure 2) by choosing the vertices of the rectangle to change, or deleted (Label 6 in Figure 2) by removing a rectangle.

The tool includes a labels bar (Label 7 in Figure 2), where different color labels are presented and can be chosen to represent different annotations (e.g., if a weather map has two storms, it should be annotated with two different labels). A checklist with the metric values is included, where the geometric objects of the weather metric can be filtered by their respective values (Label 8 in Figure 2). The annotator tool can load a panoply of maps and it is possible to choose the map to be displayed (Label 9 in Figure 2). The annotations are saved automatically to the map represented, so that it is possible to return to a previous map and load the respective annotations. The feature of navigating through a series of maps can be used to upkeep with the evolution of a meteorological phenomenon.

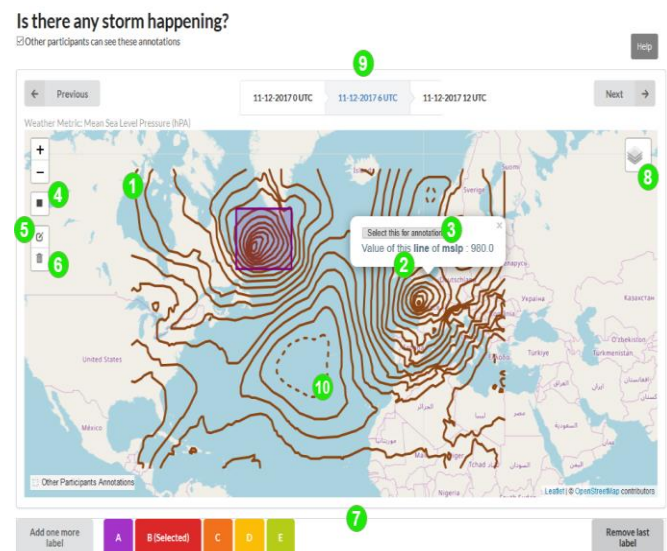


FIGURE 2. User interface of the Extreme Weather Events Annotator.

Additional annotations made by previous crowd workers can also be displayed in the annotator tool. Such annotations are represented in dashed lines (Label 10 in Figure 2) and cover the whole or part of the weather geometric object. For determining whether the other crowd worker annotation is represented, a threshold of 50% was defined, based on the weather geometric object's perimeter covered by the

<sup>5</sup> <https://www.djangoproject.com/>

<sup>6</sup> <https://matplotlib.org/>

<sup>7</sup> <https://cds.climate.copernicus.eu/>

<sup>8</sup> <https://reactjs.org/>

<sup>9</sup> <https://leafletjs.com/>

annotation. The reason for the definition of this threshold was two-fold:

- If the threshold was defined as 0%, it would allow the representation of any other crowd worker's annotation even if just covered a small portion of the weather geometric object. The consequences of this option would be the appearing of false positives on extreme weather events annotated by the other crowd workers in the interactive map;
- If the threshold was established as 100%, the annotations provided by other crowd workers would not be represented because it did not cover the whole perimeter of the weather geometric object. This would cause the appearance of many true negative extreme weather events that could be not annotated by other crowd workers.

#### D. Crowdsourcing Platform

For the presentation of crowdsourcing campaigns to the crowd workers, a platform that could host the annotator tool and integrate with the eCSAAP system was necessary. Thus, we selected the PYBOSSA<sup>10</sup> crowdsourcing platform, as it enables the implementation of a task presenter adjustable for tailored crowdsourcing campaigns, which is important to enhance the interface shown to the crowd worker. This platform comes with an API that allows its integration for managing the tasks and importing the results. In addition, PYBOSSA gives support to enrich the task presented to the crowd worker, including a built-in mechanism for passing task information to the task presenter. Furthermore, it is possible to retrieve the finished tasks results during the task run.

As shown in Figure 1, the task is added in the crowdsourcing platform by the eCSAAP system. For the implementation of the task presenter in the crowdsourcing platform, the Extreme Weather Events Annotator is used. When a task is presented to a crowd worker, the visualization data is fetched from the eCSAAP system and loaded to the interactive map of the annotator tool. Moreover, if another crowd worker has done the same task, it also includes his/her annotations so that the current crowd worker can observe the annotations made by other users.

#### IV. Research Method

In order to evaluate the influence of stigmergy for weather annotations, we conducted a study on the eCSAAP-SOF4WA. The goal of the study was to identify the work done by experts when annotating atmospheric phenomena and, more specifically, the effect of stigmergic collaboration in crowdsourcing. In this section, we describe the tasks proposed, the data collected, and our approach to data analysis.

##### A. Experimental Design

The study followed an experimental design based on the following questions:

- *Individual vs collective annotations.* Can collective annotations take advantage of individual analysis of weather data?
- *Influence of the crowd.* How does a crowd worker's annotation influence the following ones? The conditions for the evaluation are the display of: no crowd worker annotations, real crowd worker annotations, simulated positive annotations, or simulated negative annotations.

For each of the four distinct research conditions a task was created that consisted of annotating the evolution of weather storms. Users were presented with three weather maps that represent events with 6-hour intervals between them (e.g., if the first map showed time 0 UTC, the second would be an image taken at 6 UTC and the third one at 12 UTC). The task was differentiated from the other crowd worker annotations in order to observe the behaviors when the crowd worker is able to view the results of the other crowd workers, specifically studying the Bandwagon effect, which is a cognitive bias, on the stigmergic process [43]. Each type of task is presented below:

- *Task 1. Simulated negative annotations.* In this task, annotations are displayed where no storm is happening (map is incorrectly annotated). The crowd worker may be induced to follow precedent, thus marking storms where they are not occurring.
- *Task 2. No crowd worker annotations.* The crowd worker finds the weather map without any previous annotations, so there should be no influence upon the crowd worker's annotations.
- *Task 3. Simulated positive annotations.* A task is presented to the crowd worker with results that indicate annotations corresponding to extreme weather events (map is correctly annotated). This should help the crowd worker to annotate actual storm events more effectively. In this condition, real crowdsourcing annotations from other volunteers are not shown.
- *Task 4. Real crowd worker annotations.* The map presents annotations provided by other users. This condition could influence the crowd worker toward making either true or false annotations, depending on what other users have previously generated.

The traces resulting from participant contributions towards the task outcome were recorded. Such traces help us track the work on the shared output of the project and thus can be used as a valid measure of stigmergy [88]. For the case study, the objective of each task was to annotate the storms on the weather metric of Mean Sea Level Pressure (MSLP). The MSLP is typically represented in charts as lines, and each

<sup>10</sup> <https://pybossa.com/>



line has an equal pressure value [89]. An explosive cyclogenesis occurs if there is a rapid pressure drop [90]. Furthermore, many lines close to each other may indicate a rapid pressure drop. The weather map presented in the task is composed of MSLP lines.

### B. Procedure

The target participants for the study were weather experts, including climate and weather scientists and enthusiasts. Enthusiasts are a special group that includes professionals with advanced knowledge of weather phenomena (e.g., aircraft pilots) and individuals whose hobbies require similar weather knowledge, which include but are not limited to paragliding pilots, surfers, and natural life photographers. Recruitment was conducted by sending an email invitation to target groups mailing lists. The invitation contained the explanation of the study, and how to proceed to enroll in the experiment, and access the pre-task questionnaire, the crowdsourcing platform, and the post-task questionnaire.

A pre-task questionnaire was used to understand the participants' profile, including the proficiency in performing online tasks and previous working background related to the climate field. These two questions were presented on a 4-point Likert scale ranging from strong agreement to strong disagreement with each item. In the crowdsourcing platform, participants were given instructions for annotating events in a weather task which included a video with duration under 4 minutes that was displayed on the crowdsourcing platform in order to briefly explain both how the interface works when creating annotations, and how to interpret the MSLP for the identification of storms. After watching the video, the participants could start performing the task. Each task consisted of three weather maps, and participants were asked to track the evolution of storms.

The participants were instructed to annotate multiple storms that could be presented on a single weather map, and to identify the same storm in the following weather maps of that task. There was no time limit and the campaign included a step-by-step tutorial that highlighted and explained each task component which could be used as necessary. After the completion of the task, the participants filled in the post-task questionnaire described next.

### C. Assessment

The tasks consisted respectively in annotating maps of the following storms, represented in the weather metric of MSLP (these named storms occurred in the Euro-Atlantic region and drew adequate media coverage to be familiar to most participants):

- '*Ana*' Storm (10–14 December 2017)<sup>11</sup>
- '*Felix*' Storm (9-16 March 2018)<sup>12</sup>

<sup>11</sup>[https://www.eumetsat.int/website/home/Images/ImageLibrary/DAT\\_3759905.html](https://www.eumetsat.int/website/home/Images/ImageLibrary/DAT_3759905.html)

- '*Gong*' Storm (18-22 January 2013) [90]
- '*Klaus*' Storm (23-28 January 2009) [91]

Each task consisted of one storm, during its peak stage. Storms were previously identified with the support of surface analysis data from the UK Met Office<sup>13</sup>, that included the locations of the weather fronts, low- and high-pressure systems. The assessment included annotations of regions that could include storms, specified with the help of a weather expert, which analyzed this category of storms [90, 91]. The map for annotation of the storms was created to show not only the storm itself, but to include regions where the MSLP could indicate potential storms. This was included to widen the spectrum of possible annotations by the crowd workers, who could potentially identify a multitude of storms. The annotation is represented with a bounding box (an annotation made with a rectangle) covering the storm area on a map image.

The evaluation for this study focused on assessing group consensus simple majority, a quality measure in crowdsourcing [92], by generating labels from the data based on the most representative answer from the crowd. We analyzed the effects of stigmergy on the quality of weather data annotations through a comparison of annotations done with or without awareness of past contributions provided by the other participants.

Aggregation of crowd work can be done through visual annotations with effective results [93]. In this regard, Oosterman and colleagues evaluated the group consensus when labeling artwork images using bounding boxes. The aggregation method comprised three steps:

1. *Preprocessing*. The bounding boxes that were not precise were removed (area larger than 3 standard deviations of the mean value of all bounding boxes). This step is useful because one bounding box could have a high coverage of the pretended observation, but would lose quality if the proportion is too high.
2. *Clustering*. The bounding boxes were grouped into clusters based on overlap. This step directs each annotation in a cluster to the same visualization target. Clustering techniques like the K-means and the Gaussian Mixture Model were used for clustering.
3. *Aggregation*. For each cluster, the aggregation result can be obtained by making the derivation of each cluster using the maximum or the median values.

Taking into consideration that crowd workers can introduce noise to the data (bias), clustering can be a feasible path to infer ground truth values [94]. The case study

<sup>12</sup>[https://en.wikipedia.org/wiki/2017%E2%80%9318\\_European\\_winter\\_storm\\_season#Storm\\_Felix](https://en.wikipedia.org/wiki/2017%E2%80%9318_European_winter_storm_season#Storm_Felix)

<sup>13</sup><https://www.metoffice.gov.uk/weather/maps-and-charts/surface-pressure>



proposed in this article followed the methodology defined by Oosterman et al. [93] to aggregate crowd workers' results. Nonetheless, we adopted the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) technique for clustering, which is recommended for clustering spatial domains [95]. The maximum distance between points to be considered inside the clusters is a necessary metric for applying the DBSCAN method. Maximum distance was set to 3 degrees (in the geographic coordinate system) taking into consideration the wide dimension of the domain for annotations covering the entire Euro Atlantic region.

Task evaluation was measured with the ratio of area selected by the participant and the median bounding box from the group consensus clusters. The quality measure used for this evaluation was the F-score [96], which also has been applied to measure the effectiveness of annotations for evaluation of crowdsourcing tasks [29]. The F-score compares the area coverage of the annotation against the ground truth, taking a proportion of the area drawn. This metric allows us to evaluate the effectiveness and precision of annotation. Effectiveness relates to the coverage of the median cluster bounding box and precision to the proportion (e.g., if an annotation covers the whole cluster median bounding box but is a lot larger, it has less precision and, consequently, lower quality). A log with the participants' actions was kept, recording where the user clicked while completing the task.

To provide a qualitative perception, a post-task questionnaire debrief was given to participants. Six questions were asked, using the 5-point Likert scale to measure satisfaction, related with the evaluation of the task presented to them:

1. How do you rate the representation of the weather map? (User experience's satisfaction)
2. Do you feel satisfied with the annotations made? (User experience's satisfaction)
3. Did you understand the annotations of the other volunteers? (Assessment of the results provided by other users)
4. Were you influenced by the annotations of the other volunteers? (Assessment of the results provided by other users)
5. Could you observe the evolution of the weather map? (Awareness of the map)
6. How was the quality of the interactions' tools? (Evaluation of the quality of the user interface)

The post-task questionnaire also included an open-ended question so that participants could provide feedback, suggestions, or report problems they encountered. Figure 3 presents a technical flowchart that summarizes the research method of this study.

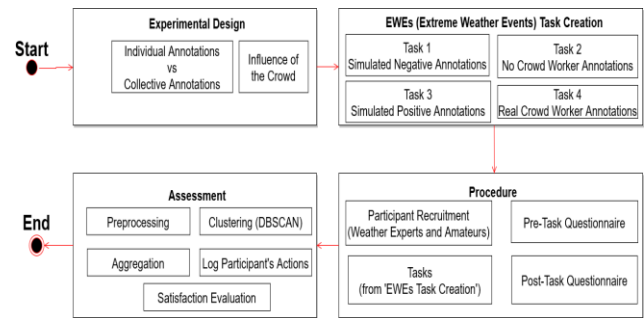


FIGURE 3. Technical flowchart of the research method.

## V. Results

In this section, we report both quantitative and qualitative results of the experiments. Furthermore, we investigate the relative importance of stigmergy and task modularity on the general outputs provided by experts.

In this study there were 13 participants, with an average age of 42.14 (SD=14.98). Twelve participants had working experience in the climate field, varying from weather experts that research atmospheric phenomena to people that assisted with seasonal forecasting. One participant was an enthusiast of the field, being interested in weather forecasting due to a related hobby (i.e., paragliding). Figure 4 presents a chart with the answers to the pre-task questionnaire related to the proficiency in performing online tasks. As it can be observed from Figure 4, most of the participants stated that they were proficient at performing online tasks.

### A. Individual vs. Collective

The assessment of the group consensus was made by aggregating the results of the participants into clusters. Figure 5 presents the distribution of the values associated with the task performance. The clustering had lower dispersion and worst results (according to the F1 scores) in Task 2 (where no other crowd workers results were shown). As the crowd workers could not see each other's results, their annotations had more variation. Task 1 (Simulated Negative Results) had reasonable values with a reasonable dispersion. Task 3 (Simulated Positive Results) had satisfactory precise dispersion, obtaining good results in the contribution of clusters. Moreover, Task 4 presented the real results of the other crowd workers to each participant and, although it obtained the maximum value of the F-score, it also had the largest range in values among the four tasks.

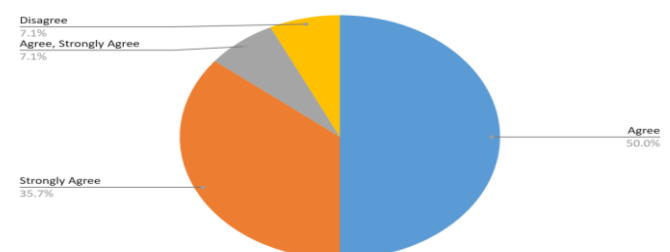


FIGURE 4. Pre-task questionnaire answers to the item of proficiency in performing online tasks.

The task duration and the number of clicks increased progressively through the tasks, with a more notable highlight in the Task 4. The presentation of real results from other crowd workers, could suggest a more difficult situation for the participant to complete the task, and also indicates that participants attempted to complete the task to the best of their ability, since each task variation incorporates more complexity.

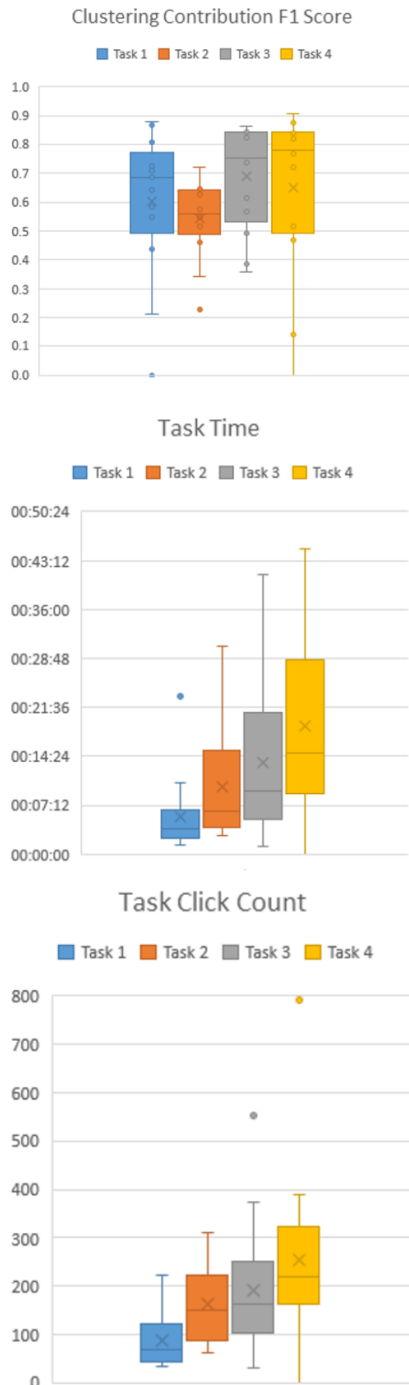


FIGURE 5. Box plot of the Clustering Contribution measured with the F1 Score (left), the Task Duration (middle) and Task Click Count (right).

Table 1 presents the individual results in this case study, showing the evolution of each participant during the task execution. The participant that was not a weather expert (P1) had the worst results of the group of participants, but still contributed in each task with at least one annotation per cluster, without ever drawing an annotation outside the clusters.

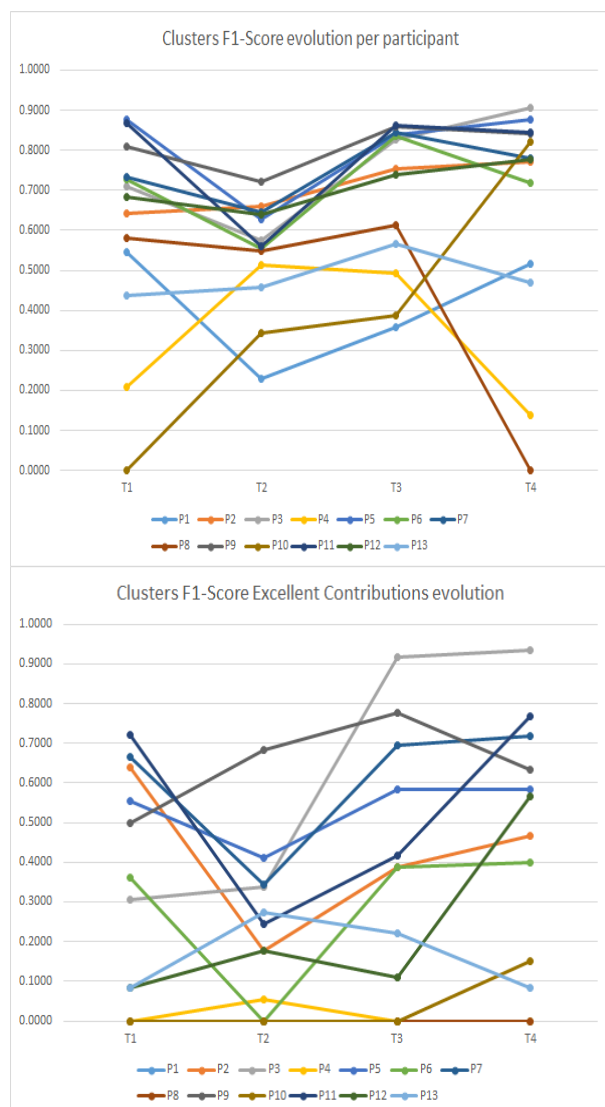
The individual results show that the users who contributed for forming clusters had good proficiency using the system (as measured by total click count and task time) (P3, P6, P7, P9, P11). These results were accompanied by good F-scores and a significant appearance of annotations inside the clusters. Extremely fast task performers had bad results (P8, P10), which could be due to not being interested in participating in the case study.

TABLE 1. Overview of individual results on the evolution of participants over the tasks.

Participant id	Average Ratio F1 Dice Score Per Cluster				Total Click Count				Task Time				Number Annotations Outside Cluster				Ratio Annotations Inside Cluster				Ratio Annotations Inside Cluster with F1 Dice Accuracy >= 80%			
	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4
P1	0.5459	0.2290	0.5388	0.5105	222	299	375	399	00:20:11	00:30:38	00:41:07	00:44:51	0.0000	0	0.0000	0.0000	0.5554	0.7056	0.3103	0.2225	0.0000	0.0000	0.0000	0.0000
P2	0.6416	0.6194	0.7028	0.7702	43	182	207	242	00:03:01	00:10:48	00:12:41	00:14:51	0.0000	1.3333	0.3333	0.0000	1.0000	0.7056	0.5037	0.3847	0.6389	0.1779	0.0000	0.4867
P3	0.7100	0.5742	0.8225	0.7667	45	95	181	270	00:03:03	00:06:23	00:11:58	00:18:02	0.0000	0.6667	1.0000	0.3333	1.0000	0.8222	1.0000	1.0000	0.3016	0.3333	0.5047	0.9383
P4	0.2905	0.5340	0.4011	0.1393	44	79	122	204	00:05:57	00:05:30	00:21:27	00:26:21	0.0000	0.3333	0.0000	0.0000	0.3444	0.2056	0.4387	0.3333	0.0000	0.0000	0.0000	0.0000
P5	0.7101	0.6263	0.8396	0.8755	111	209	294	378	00:05:26	00:15:30	00:20:12	00:24:04	0.0000	0.0000	0.0000	0.0000	0.7222	0.8222	0.5037	0.7500	0.5556	0.4111	0.5833	0.5833
P6	0.7528	0.5540	0.8543	0.7255	37	84	113	175	00:05:26	00:05:45	00:04:50	00:07:26	0.3333	0	0.0000	0.3333	0.7222	0.2056	0.6111	0.5000	0.3631	0.0000	0.3409	0.4000
P7	0.7528	0.6400	0.8840	0.7810	93	207	314	238	00:04:31	00:14:55	00:04:19	00:10:22	0.0000	0.6667	0.3333	0.0000	0.0000	0.7500	1.0000	0.5000	0.3444	0.0000	0.3444	0.7500
P8	0.3819	0.5400	0.6124	0.7010	138	121	108	101	00:05:47	00:05:00	00:06:22	00:06:22	0.0000	0	0.0000	0.0000	0.5556	0.3500	0.2000	0.0000	0.0000	0.0000	0.0000	0.0000
P9	0.9394	0.7121	0.8596	0.8845	70	111	108	144	00:01:57	00:08:55	00:01:52	00:01:52	0.0000	0.6667	0.3333	0.0000	0.5556	0.8222	1.0000	1.0000	0.5000	0.6667	0.7778	0.6111
P10	0.3000	0.3400	0.3800	0.3100	76	95	10	14	00:01:52	00:04:00	00:01:07	00:01:34	0.0000	0.3333	0.0000	0.3333	0.0000	0.2056	0.0000	0.2000	0.0000	0.0000	0.0000	0.0000
P11	0.8678	0.5638	0.8638	0.8654	72	121	142	214	00:06:28	00:06:24	00:09:19	00:13:06	0.0000	0	0.0000	0.0000	0.5000	0.5000	0.5000	0.5000	0.7222	0.5000	0.6111	0.7944
P12	0.8884	0.6188	0.7379	0.7782	59	89	162	224	00:02:41	00:04:02	00:07:14	00:10:46	0.0000	0	0.0000	0.0000	0.5000	0.5000	0.5000	0.5000	0.6667	0.5000	0.5000	0.5000
P13	0.6380	0.4030	0.5864	0.4074	109	102	104	70	00:08:25	00:16:24	00:28:51	00:42:34	0.6667	2.0000	2.0000	0.0000	0.8000	0.6167	1.0000	0.5000	0.6667	0.7222	0.7222	0.6667

## B. Influence of the Crowd

Each participant was evaluated in order to assess whether or not presenting content that showed annotations influenced their performance. Figure 6 depicts two charts related to the clusters' contributions of the participants. The average F-score results were stable for most users, with a slight increase on Task 4. However, some users performed much worse (P4, P8). In particular, P4 had previously provided poor contributions, seldom contributing with excellent annotations to the clusters, and was the only participant with the false positive condition in Task 1. P8, although presenting good results, did not submit the annotations in the last task. Most of the participants increased the quality of annotations on Task 3 but either increased or worsened the results on Task 4. This could be due to the fact that Task 3 simulated positive (accurate) crowd workers results, which helped the improvement of the contributions. When subsequently presented with Task 4, the participants could encounter either accurate or misleading annotations from the real crowd workers, which could confuse the participant when choosing regions where the clusters would be present.



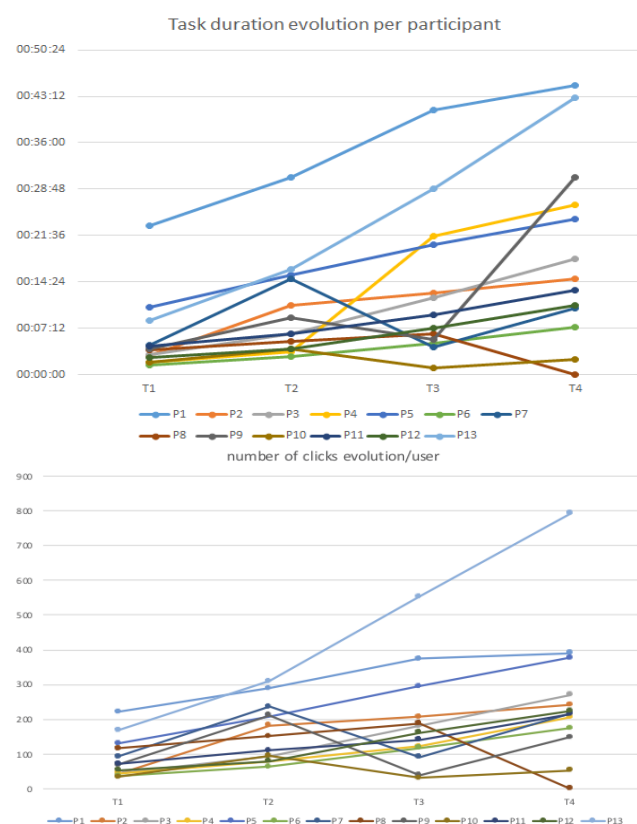
**FIGURE 6.** Line Chart of the Average Clusters F-Score evolution (top) and Ratio of Excellent (>0.8 F-Score) Contributions evolution (bottom).

Figure 7 presents Line charts showing the evolution of each participant's performance during the tasks. Both task duration and number of clicks increased progressively through the completion of the tasks, reaching maximum results in Task 4. This may be because in Task 1 the participants were presented with negative results, and most of them figured out that such region would not have a storm. In Task 2, there were no other crowd workers results, so it did not affect the participant time or number of clicks but, as shown earlier, also did not increase the quality of contributions. In Task 3, the positive results suggest that the user takes more time and actions to perform the annotations with good contributions. Last but not least, Task 4 had the highest number of duration and clicks, with several participants taking more time and actions for adding final annotations, which suggests that the task required additional

processing to identify if the annotations present were accurate or misleading.

### C. Satisfaction Evaluation

At the end of the study, participants had to fill a post-task questionnaire with questions about their experience. Two participants did not answer the post-task questionnaire (P3, P9). They rated the quality of the weather map as neutral (9.1%), satisfied (63.6%) and very satisfied (27.3%). When asked about keeping up with the evolution of the weather maps (each task had three weather maps chronologically related), 54.5% replied that they were satisfied and 45.5% were very satisfied. When asked about if they feel satisfied, most of the participants gave positive feedback (Unsatisfied - 9.1%, Neutral - 45.5%, Satisfied - 27.3%, Very Satisfied - 18.2%). The question of how the quality of the interactions' tools obtained good results.



**FIGURE 7.** Line chart of the evolution on the user task duration (left) and number of clicks (right).

Regarding the responses to the question about understanding the annotations of other volunteers, most participants seemed to agree that they understood the annotations (54%) and 27% replied with "neutral" which would suggest that they either understood some but not all annotations, or they were not sure whether or not they understood the annotations correctly. This suggests that the participants did not understand and neither felt influenced by

other crowd workers' results. However, the results presented previously point to the participant being influenced, either positive or negatively by each other's results.

In the post-task questionnaire, the last item was open-ended, meaning that the participants could write their opinion and feedback regarding their participation. The open-ended item was optional and was answered by six participants (P1, P5, P6, P7, P10, P12). Among these comments, participants praised the accessibility and utility of the tools, including one remark about applying this tool in the professional context:

*"Very interactive and easily accessible platform" (P12)*

*"I realized that I can apply that tool in my classes" (P7)*

One participant (P5) highlighted that a better example of a storm should be given. The goal of this study however, is to get annotations from a wide spectrum of extreme weather events severity, which can be light storms to cyclones. Other participants suggested improvements, suggesting that it could be inserted open-ended text to the annotations made, so that the other crowd workers could see. Related with the interactivity, one participant said that the interactivity could highlight better the values to annotate but should be done carefully.

*"From the tutorial I had the impression I could write text related to rectangle, but I could not find the tool if it was available. Note I did not find the other users annotations – I clicked at some point in the bottom (in one of the examples and there was no output)" (P6)*

*"[...] require some improvement through more interaction. [...] while annoying, it may highlight values. This will be more useful, thanks." (P10)*

A participant that demonstrated interest in the climate field as an enthusiast (P1) felt some difficulties in using the annotator due to a lack of experience with online tools. The support of the Help Guide Tour feature was identified as a useful resource.

*"[...] the 'Help' menu helped a lot to understand everything that the remaining tasks occurred with normal progress." (P1)*

The challenges faced were also overcome with the participant's curiosity on the climate field and the perception that the tasks were interesting.

*"[...] I think that the tasks were interesting, but the knowledge about Pressure Lines is not enough to produce more concise results." (P1)*

## VI. Discussion

Participants in the study were quite a homogeneous sample composed primarily of scientists (92%); the community of enthusiasts was represented by one participant (8%). This expert crowd has a reduced size, limiting the usage of statistical methods, but allowing a qualitative analysis. Participants revealed that they can easily perform online tasks (95% agree/fully agree). Users' experience with the platform was generally positive, with high levels of satisfaction. Participants' perceptions of the utility of other users' annotations were mixed, and future work could

examine this aspect of the interaction design more closely, as the results also indicated that the visibility and sources of prior annotations impacted their work.

### A. Stigmergic effects on annotation

The scope of the study considers individual and group behaviors and the influence of the crowd as an intrinsic coordination strategy (stigmergic effect). The analysis of individual behavior enables perception of each participant's work consistency and their event annotation behavior. The results showed a high individual consistency in the density of users' annotations as well as their granularity, but with variability between participants. The individual differences in their understanding of the semantic domain of the concept of "extreme atmospheric phenomena" indicates a lack of common ground that could impact the quality of aggregate annotations: for some experts, a storm is any extreme weather phenomenon, however, others had a more restricted understanding, considering only hurricanes to qualify as extreme phenomena. This diversity is also reflected in the precision and detail with which the phenomena were analyzed by each expert, which likely relates to their specific knowledge from their climate science backgrounds, and the only enthusiast in the group did not reach the expected results.

However, the quality of the results reveals that some scientists also stray from the pattern when annotating phenomena. When we examined the number of interactions (mouse clicks) and time of each interaction, a particular situation was identified: quick answers, with a reduced number of interactions, very low precision and data quality, that may reflect the participant's low motivation to perform the task or other factors such as time constraints or environmental distractions. Further research, perhaps using a think-aloud protocol, would help uncover the reasons for poor performance. In general, the most accurate data resulted from an intermediate interaction density and equivalent task duration. These results can be explained by the users' ability with the platform, familiarity with technology, and background related to the tasks (e.g., prior experiences with the specific weather events in the images presented.) Nevertheless, there was no direct relationship between the quality of the work completed and the participants' satisfaction with platform based on the post-questionnaires.

The individual progression of each participant in terms of time spent and density of interactions follows an increasing trend between the tasks. It is not clear why there is an increase in the average time dedicated to each task. This may be related to the increasing complexity of the phenomena presented. A note that there is no evidence of a relationship between the time spent and the complexity of the interface and the density of the annotations, since task 1 includes simulations of previous annotations whereas task 2 does not contain any annotations, and this decrease is not reflected in this passage of the average interaction time.



If the analysis of the time spent on each task is inconclusive, the same is not true for the quality and accuracy of the data. The analysis of the clustering variance of contributions reveals that there is a decrease in task 2, that is, in the only task in which there were no previous annotations (simulated or real). Therefore, there is strong evidence for the influence of previous annotations on the quality of the obtained annotation data. This result is visible, either in an individual analysis or in a grouped analysis of the results. Another interesting aspect is that there is no impact of the nature (real or simulated) of previous annotations, regardless if they are true or false. The results show that the existence of previous annotations denotes a certain “comfort” of the participants and consequently better results, which is aligned with the underlying assumptions of the Bandwagon effect [43]. However, they do not demonstrate this perception in the post-task questionnaires, revealing that there was no significant influence when displaying prior annotations. This lack of awareness suggests that this interface feature can be suitable for the stigmergy process without causing entropy to the user experience.

Other important aspects to keep in mind are the convergence towards fine granularity of the annotation, not the grouping of phenomena. The initial analysis indicates that the annotations of the phenomena were broad annotations. However, the convergence of expert crowd annotation shows that a fine annotation granularity showing the individual phenomena is preferred instead of grouping the phenomena. This result is very interesting and it can be inferred from here that the phenomena must be analyzed in their individual characteristics and only then can the interactions between different phenomena be understood, since human perception tends to ungroup phenomena. However, the question remains how and when to group the phenomena. Experienced researchers and experts in the reported phenomena initially identified the cluster. This is a limitation and it will be necessary to do new tasks to assess the researchers' ability to group phenomena.

Stigmergy has emerged as a compelling and innovative coordination mechanism in crowdsourcing, demonstrating its potential in the domain of weather forecasting. Our study showcases the power of stigmergic collaboration among weather experts in annotating and predicting extreme weather events, with a particular emphasis on storm forecasting. Stigmergy offers several unique advantages over other forms of coordination. Unlike traditional hierarchical methods that rely on explicit communication and coordination among participants, stigmergy leverages indirect communication through the shared environment, enabling crowd workers to build upon each other's annotations autonomously. This self-organization aspect fosters a dynamic and decentralized collaboration, where participants respond to the collective intelligence of the crowd, leading to emergent and accurate predictions. Furthermore, stigmergy harnesses the collective expertise of diverse weather experts, allowing for a more

comprehensive analysis of weather data and mitigating individual biases. The results of our study demonstrate that stigmergic collaboration enhances the quality and effectiveness of weather annotations, making it a promising coordination mechanism for weather forecasting research [44, 67].

While stigmergy shows great promise, it is essential to acknowledge the role of other coordination mechanisms in crowdsourcing, as they each have their strengths and limitations. For instance, centralized coordination methods, such as expert-led annotation platforms, provide direct oversight and control over the crowdsourcing process, ensuring high-quality annotations but potentially limiting scalability due to resource constraints. On the other hand, decentralized coordination mechanisms, like peer-to-peer collaboration, offer flexibility and scalability, but may struggle with quality control and coordination among participants. Comparative studies between stigmergy and these other approaches can shed light on the trade-offs and applicability in different contexts. Moreover, combining different coordination mechanisms in a hybrid approach may offer synergistic benefits, leveraging the strengths of each mechanism while mitigating their weaknesses. Future research could explore such hybrid models and investigate how stigmergy can complement and enhance other coordination methods in diverse crowdsourcing applications, not only in weather forecasting but also in various scientific and societal domains [55, 61, 68, 69].

### **B. Current Challenges and Limitations**

A major limitation of this research is that the experiment was conducted on a small scale with just 13 participants, 12 experts and 1 enthusiast (i.e., a paraglider with more than 30 years of expertise on assessing weather phenomena for flying in an optimal way). The experiment targeted a very specific group, with very particular capabilities which restricted participant recruitment. The limited number of participants constrained the results methodology as statistical methods would be inappropriate. However, this limitation was an opportunity to explore a different approach for gathering collective knowledge from a largely homogeneous, expert crowd. Further studies are needed to evaluate the behavior of enthusiasts and the impacts on the data quality of their collaboration in the expert crowds, as this study was unable to do so. On the other hand, the small sample size supported an in-depth review of participant behaviors during the task execution. The small sample size based on experts is validated in crowdsourcing settings. In line with this, Kazai and Zitouni [97] concluded that using a sample of fewer than 16 experts in two relevance label tasks significantly enhanced the quality of the data collected. Furthermore, in the HCI field, having a small sample such as the five-user assumption can be used as an optimal benchmark to identify interaction problems [98-100].

The research also did not include a usability study of the interface. The system was co-created with experts using a participatory design development approach [85], according to their requirements and following their visualization golden rules. However, we are unable to evaluate any potential relationship between the usability of the platform and the data quality. The implication of this limitation is that real contributions may overload the visualization tool. In addition, this study showed that previous annotations contribute to enhancing data quality, but does not allow us to conclude what kind of annotations (real vs simulated).

The expertise of the participants provided high data quality and the preliminary results demonstrate some open research opportunities in the frontier between IS and meteorology and atmospheric sciences. As part of an exploratory project, this study establishes a research roadmap to explore the open challenges in the use of human computation [22] and artificial intelligence [20] to study extreme atmospheric phenomena. Further research is required in order to evaluate the impacts of real and simulated annotations, as well as simulated true and false annotations. Moreover, studying crowd processes across different levels of expertise would be useful further research.

### **C. Implications for Design**

#### **1) USER INTERFACE DESIGN**

Driven by the increasing availability of data produced through massive collaboration efforts, Gadiraju et al. [91] investigated the effect of User Interface (UI) design elements in microtask work environments and concluded that bad design can affect the overall outputs generated by crowd workers. The eCSAAP-SOF4WA Extreme Weather Event Annotator can support implicit coordination mechanisms, by presenting previous annotations of extreme weather events to crowd workers, but this potentially useful feature must be balanced against task complexity resulting from large numbers of annotations. These strategies reveal, already used in [94], were applied to process the results of the current study, enhancing the collective knowledge. The integration of human factors such as expertise and experience obtained from a collective of weather specialists can enhance the precision of meteorological prediction systems. Utilizing the inherent collaboration of expert crowds in semantic labeling can potentially augment readiness for severe meteorological phenomena and mitigate their consequences. A primary consideration in user interface design is achieving a balance in the presentation of results from other crowd workers. In the context of interactive web maps, this can be accomplished through the modification of geometric object styles and the implementation of dynamic pop-ups, as demonstrated in this study. The former involves the use of dashed lines to represent Extreme Weather Events (EWEs) annotated by other weather experts, which influenced their decision-making. The latter involves the inclusion of additional information in pop-ups activated by clicking on

geometric objects. These two features facilitate the display of adaptive information that can aid decision-making without overwhelming the crowd worker.

#### **2) MOTIVATION FACTORS**

Concerning the motivational aspects underlying the proposed solution, a design principle commonly used to enhance the quality of the data relies on providing continuous feedback [101]. From a holistic perspective, this study demonstrates that prior annotations can act as a critical factor for improving data quality. However, the mechanisms through which the annotations serve to improve quality are not clear. Expert crowds participation do not follow traditional behavioral patterns in crowdsourcing settings (e.g., [102]), so specific strategies are required for motivating expert crowds, but the unexpected behavior (i.e., low quality annotations) for some experts in this study suggests a need for further evaluation of the reasons for this result, which may be related to motivation [103]. Aitamurto and Saldivar demonstrated that motivation factors weakened and the crowd disengaged when crowdsourcing reached a saturation point. Sustaining participation is one of the core challenges for crowdsourcing systems, which requires knowledge of the crowd's expectations and motivations.

#### **3) EXPERT CROWDSOURCING AND COMMUNITY OF EFFORT**

Volunteer contributors working worldwide in the face of global climate change constitute a great example of a community of effort (a community formed in pursuit of a common goal). In such scenarios, no central organization coordinates; rather, the scientists “collaborate in ad hoc ways, and are conscious of contributing toward a shared purpose” [104]. The eCSAAP follows this strategy by creating ways to involve a broad community in understanding of extreme atmospheric phenomena. Expert crowdsourcing is a first step in this strategy, and extending the expert crowds to include weather enthusiasts would be a next step as enthusiast communities have particular characteristics that could be an asset for annotation of extreme phenomena. However, this study was not able to adequately assess the potential value of combining the work of experts and enthusiasts, so further research will be needed to evaluate the participation of a broader community in the study of extreme atmospheric phenomena.

## **VII. Conclusions and Future Work**

Crowdsourcing represents an opportunity for novel research and practice in climate sciences. Extreme atmospheric phenomena are becoming more frequent and require alternative approaches for both the study of the events and predicting them. This research employed expert crowdsourcing to annotate extreme atmospheric phenomena using a technological platform developed with a participatory co-created design that engaged the climate science research

community in the development process. The resulting annotations, aggregated with a clustering approach, showed that new knowledge can arise from the participation of the expert. However, the engagement of expert crowd workers requires specific strategies adapted to specific requirements of this community. Through the utilization of human factors such as expertise and practical experience, we can optimize our readiness for severe weather phenomena and mitigate their consequences. Our research outcomes hold significant implications for enhancing weather prediction systems and bolstering disaster preparedness measures. The seamless integration of proficient crowd workers, equipped with a background in climate science research, can be facilitated via a tailored web-based interactive map and the application of conventional crowdsourcing techniques, such as annotation clustering. Extending the expert crowd to climate enthusiasts revealed the same engaging problems and, therefore, the size of the dataset was very limited and constrained the results of the study. Nonetheless, the results of this work identify interesting potential and future directions for the application of crowd computing in climate science research.

## ACKNOWLEDGMENT

This work is financed by the Portuguese Foundation for Science and Technology (Fundação para a Ciência e a Tecnologia – FCT) with the research grant SFRH/BD/148991/2019. The authors also acknowledge support from the European Social Fund under the scope of North Portugal Regional Operational Programme.

## REFERENCES

1. Agrawal, A.A., D.D. Ackerly, F. Adler, A.E. Arnold, C. Cáceres, D.F. Doak, E. Post, P.J. Hudson, J. Maron, and K.A. Mooney. *Filling key gaps in population and community ecology*. *Frontiers in Ecology and the Environment*, 2007. **5**(3): p. 145-152.
2. Jentsch, A. and C. Beierkuhnlein. *Research frontiers in climate change: effects of extreme meteorological events on ecosystems*. *Comptes Rendus Geoscience*, 2008. **340**(9-10): p. 621-628.
3. Auffhammer, M. *Quantifying economic damages from climate change*. *Journal of Economic Perspectives*, 2018. **32**(4): p. 33-52.
4. Mitchell, D., C. Heaviside, S. Vardoulakis, C. Huntingford, G. Masato, B.P. Guilloid, P. Frumhoff, A. Bowery, D. Wallom, and M. Allen. *Attributing human mortality during extreme heat waves to anthropogenic climate change*. *Environmental Research Letters*, 2016. **11**(7): p. 074006.
5. Smith, K., G. Kent, D. Slater, G. Plank, M. Williams, M. McCarthy, F. Vernon, N. Driscoll, H.-W. Braun, and R. Anderson. *Integrated multi-hazard regional networks: Earthquake warning/response, wildfire detection/response, and extreme weather tracking*. *Applied Geology in California: Association of Environmental and Engineering Geologists (AEG) Special Publication*, 2016. **26**: p. 599-612.
6. Yore, R. and J.F. Walker. *Early warning systems and evacuation: rare and extreme versus frequent and small-scale tropical cyclones in the Philippines and Dominica*. *Disasters*, 2021. **45**(3): p. 691-716.
7. Gaffigan, M. *Climate Resilience: A Strategic Investment Approach for High-Priority Projects Could Help Target Federal Resources*, Statement of Mark Gaffigan, Managing Director, Natural Resources and Environment, Testimony Before the Select Committee on the Climate Crisis, House of Representatives. in *United States. Government Accountability Office*. 2019. United States. Government Accountability Office.
8. Palen, L. and K.M. Anderson. *Crisis informatics—New data for extraordinary times*. *Science*, 2016. **353**(6296): p. 224-225.
9. Soden, R. and L. Palen. *Informing crisis: Expanding critical perspectives in crisis informatics*. *Proceedings of the ACM on human-computer interaction*, 2018. **2**(CSCW): p. 1-22.
10. Soden, R. and A. Lord. *Mapping silences, reconfiguring loss: Practices of damage assessment & repair in post-earthquake Nepal*. *Proceedings of the ACM on Human-Computer Interaction*, 2018. **2**(CSCW): p. 1-21.
11. Slonosky, V. and R. Sieber. *Building a Traceable and Sustainable Historical Climate Database: Interdisciplinarity and DRAW*. *Patterns*, 2020. **1**(1): p. 100012.
12. See, L., A. Comber, C. Salk, S. Fritz, M. Van Der Velde, C. Perger, C. Schill, I. McCallum, F. Kraxner, and M. Obersteiner. *Comparing the quality of crowdsourced data contributed by expert and non-experts*. *PloS one*, 2013. **8**(7): p. e69958.
13. Hutt, H., R. Everson, M. Grant, J. Love, and G. Littlejohn. *How clumpy is my image? Evaluating crowdsourced annotation tasks*. in *2013 13th UK Workshop on Computational Intelligence (UKCI)*. 2013. IEEE.
14. Saltz, J.S., R. Heckman, K. Crowston, S. You, and Y. Hegde. *Helping Data Science Students Develop Task Modularity*. in *HICSS*. 2019.
15. Moayedikia, A., K.-L. Ong, Y.L. Boo, and W.G.S. Yeoh. *Task assignment in microtask crowdsourcing platforms using learning automata*. *Engineering Applications of Artificial Intelligence*, 2018. **74**: p. 212-225.
16. Ren, X., X. Li, K. Ren, J. Song, Z. Xu, K. Deng, and X. Wang. *Deep learning-based weather prediction: a survey*. *Big Data Research*, 2021. **23**: p. 100178.
17. Tang, L., J. Li, H. Du, L. Li, J. Wu, and S. Wang. *Big data in forecasting research: a literature review*. *Big Data Research*, 2022. **27**: p. 100289.
18. Diffenbaugh, N.S. *Verification of extreme event attribution: Using out-of-sample observations to assess changes in probabilities of unprecedented events*. *Science Advances*, 2020. **6**(12): p. eaay2368.
19. Niforatos, E., A. Vourvopoulos, and M. Langheinrich. *Understanding the potential of human-machine crowdsourcing for weather data*. *International Journal of Human-Computer Studies*, 2017. **102**: p. 54-68.
20. Nushi, B., E. Kamar, and E. Horvitz. *Towards accountable ai: Hybrid human-machine analyses for characterizing system failure*. in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. 2018.
21. Kim, S. and L. P. Robert Jr. *Crowdsourcing coordination: A review and research agenda for crowdsourcing coordination used for macro-tasks*. *Macrotask Crowdsourcing*, 2019: p. 17-43.
22. Michelucci, P. and J. L. Dickinson. *The power of crowds*. *Science*, 2016. **351**(6268): p. 32-33.
23. Muller, C., L. Chapman, S. Johnston, C. Kidd, S. Illingworth, G. Foody, A. Overeem, and R. Leigh. *Crowdsourcing for climate and atmospheric sciences: current status and future potential*. *International Journal of Climatology*, 2015. **35**(11): p. 3185-3203.
24. Wechsler, D. *Crowdsourcing as a method of transdisciplinary research—Tapping the full potential of participants*. *Futures*, 2014. **60**: p. 14-22.
25. Gomes-Pereira, J. N., V. Auger, K. Beisiegel, R. Benjamin, M. Bergmann, D. Bowden, P. Buhl-Mortensen, F.C. De Leo, G. Dionísio, and J.M. Durden. *Current and future trends in marine image annotation software*. *Progress in Oceanography*, 2016. **149**: p. 106-120.
26. Bolici, F., J. Howison, and K. Crowston. *Coordination without discussion? Socio-technical congruence and Stigmergy in Free and Open Source Software projects*. in *Socio-Technical*



- Congruence Workshop in conj Intl Conf on Software Engineering, Vancouver, Canada. 2009.*
27. Crowston, K., E. Mitchell, and C. Østerlund. *Coordinating advanced crowd work: extending citizen science*. Citizen Science: Theory and Practice, 2019. **4**(1).
28. Shirado, H., F.W. Crawford, and N.A. Christakis. *Collective communication and behaviour in response to uncertain 'Danger' in network experiments*. Proceedings of the Royal Society A, 2020. **476**(2237): p. 20190685.
29. Cartwright, M., A. Seals, J. Salamon, A. Williams, S. Mikloska, D. MacConnell, E. Law, J.P. Bello, and O. Nov. *Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations*. Proceedings of the ACM on Human-Computer Interaction, 2017. **1**(CSCW): p. 1-21.
30. Crowston, K. and A. Rezgüi. *Effects of stigmergic and explicit coordination on Wikipedia article quality*. in *Proceedings of the Annual Hawaii International Conference on System Sciences*. 2020.
31. Gloag, E.S., M.A. Javed, H. Wang, M.L. Gee, S.A. Wade, L. Turnbull, and C.B. Whitchurch. *Stigmergy: a key driver of self-organization in bacterial biofilms*. communicative & integrative Biology, 2013. **6**(6): p. 11541-6.
32. Suchman, L.A. *Plans and situated actions: The problem of human-machine communication*. 1987: Cambridge university press.
33. Xia, H., C. Østerlund, B. McKernan, J. Folkestad, P. Rossini, O. Boichak, J. Robinson, K. Kenski, R. Myers, and B. Clegg. *TRACE: A stigmergic crowdsourcing platform for intelligence analysis*. in *Proceedings of the 52nd Hawaii International Conference on System Sciences*. 2019.
34. Lewis, T.G. *Cognitive stigmergy: A study of emergence in small-group social networks*. Cognitive Systems Research, 2013. **21**: p. 7-21.
35. Heylighen, F. *Collective Intelligence and its Implementation on the Web: algorithms to develop a collective mental map*. Computational & Mathematical Organization Theory, 1999. **5**(3): p. 253-280.
36. Heylighen, F. *Why is Open Access Development so Successful? Stigmergic organization and the economics of information*. arXiv preprint cs/0612071, 2006.
37. Heylighen, F. *Stigmergy as a universal coordination mechanism I: Definition and components*. Cognitive Systems Research, 2016. **38**: p. 4-13.
38. Haselton, M.G., D. Nettle, and D.R. Murray. *The evolution of cognitive bias*. The handbook of evolutionary psychology, 2015: p. 1-20.
39. Eickhoff, C. *Cognitive biases in crowdsourcing*. in *Proceedings of the eleventh ACM international conference on web search and data mining*. 2018.
40. Kruger, J. and D. Dunning. *Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments*. Journal of personality and social psychology, 1999. **77**(6): p. 1121.
41. Saab, F., I.H. Elhajj, A. Kayssi, and A. Chehab. *Modelling Cognitive Bias in Crowdsourcing Systems*. Cognitive Systems Research, 2019. **58**: p. 1-18.
42. Gadiraju, U., B. Fetahu, R. Kawase, P. Siehndel, and S. Dietze. *Using Worker Self-Assessments for Competence-Based Pre-Selection in Crowdsourcing Microtasks*. ACM Trans. Comput.-Hum. Interact., 2017. **24**(4): p. Article 30.
43. Bikhchandani, S., D. Hirshleifer, and I. Welch. *A theory of fads, fashion, custom, and cultural change as informational cascades*. Journal of political Economy, 1992. **100**(5): p. 992-1026.
44. Franzoni, C. and H. Sauermann. *Crowd science: The organization of scientific research in open collaborative projects*. Research policy, 2014. **43**(1): p. 1-20.
45. Cohn, J.P. *Citizen science: Can volunteers do real research?* BioScience, 2008. **58**(3): p. 192-197.
46. Haklay, M. *Citizen science and volunteered geographic information: Overview and typology of participation*. Crowdsourcing geographic knowledge, 2013: p. 105-122.
47. Kullenberg, C. and D. Kasperowski. *What is citizen science?—A scientometric meta-analysis*. PloS one, 2016. **11**(1): p. e0147152.
48. Piccolo, L., M. Fernández, H. Alani, A. Scharl, M. Föls, and D. Herring. *Climate change engagement: results of a multi-task game with a purpose*. in *tenth international AAAI conference on web and social media*. 2016.
49. Staffebach, M., P. Sempolinski, T. Kijewski-Correa, D. Thain, D. Wei, A. Kareem, and G. Madey. *Lessons learned from crowdsourcing complex engineering tasks*. PloS one, 2015. **10**(9): p. e0134978.
50. Amer-Yahia, S. and S. Basu Roy. *From Complex Object Exploration to Complex Crowdsourcing*. in *Proceedings of the 24th International Conference on World Wide Web*. 2015.
51. Cui, Q., S. Wang, J. Wang, Y. Hu, Q. Wang, and M. Li. *Multi-Objective Crowd Worker Selection in Crowdsourced Testing*. in *SEKE*. 2017.
52. Cullina, E., K. Conboy, and L. Morgan. *Choosing the right crowd: An iterative process for crowd specification in crowdsourcing initiatives*. in *2016 49th Hawaii International Conference on System Sciences (HICSS)*. 2016. IEEE.
53. Hirth, M., T. Hoffeld, and P. Tran-Gia. *Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms*. Mathematical and Computer Modelling, 2013. **57**(11-12): p. 2918-2932.
54. De Beer, J., I.P. McCarthy, A. Soliman, and E. Treen. *Click here to agree: Managing intellectual property when crowdsourcing solutions*. Business Horizons, 2017. **60**(2): p. 207-217.
55. de La Robertie, B., Y. Pitarch, and O. Teste. *Measuring article quality in wikipedia using the collaboration network*. in *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2015. IEEE.
56. Casarin, J., N. Pacquerdiaud, and D. Bechmann. *Umi3d: A unity3d toolbox to support cscw systems properties in generic 3d user interfaces*. Proceedings of the ACM on Human-Computer Interaction, 2018. **2**(CSCW): p. 1-20.
57. Lukyanenko, R., J. Parsons, and Y. Wiersma. *The impact of conceptual modeling on dataset completeness: A field experiment*. in *ICIS 2014*. 2014.
58. Lukyanenko, R., J. Parsons, and Y.F. Wiersma. *The IQ of the crowd: Understanding and improving information quality in structured user-generated content*. Information Systems Research, 2014. **25**(4): p. 669-689.
59. Irshad, H., L. Montaser-Kouhsari, G. Waltz, O. Bucur, J. Nowak, F. Dong, N.W. Knoblauch, and A.H. Beck. *Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: evaluating experts, automated methods, and the crowd*. in *Pacific symposium on biocomputing Co-chairs*. 2014. World Scientific.
60. Bonney, R., J.L. Shirk, T.B. Phillips, A. Wiggins, H.L. Ballard, A.J. Miller-Rushing, and J.K. Parrish. *Next steps for citizen science*. Science, 2014. **343**(6178): p. 1436-1437.
61. O'Leary, M.A., K. Alphonse, A.H. Mariangeles, D. Cavaliere, A. Cirranello, T.G. Dietterich, M. Julius, S. Kaufman, E. Law, and M. Passarotti. *Crowds replicate performance of scientific experts scoring phylogenetic matrices of phenotypes*. Systematic Biology, 2018. **67**(1): p. 49-60.
62. Warby, S.C., S.L. Wendt, P. Welinder, E.G. Munk, O. Carrillo, H.B. Sorensen, P. Jennum, P.E. Peppard, P. Perona, and E. Mignot. *Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods*. Nature methods, 2014. **11**(4): p. 385-392.
63. Nguyen, A.T., B.C. Wallace, and M. Lease. *Combining crowd and expert labels using decision theoretic active learning*. in *Third AAAI conference on human computation and crowdsourcing*. 2015.
64. Law, E., K.Z. Gajos, A. Wiggins, M.L. Gray, and A. Williams. *Crowdsourcing as a tool for research: Implications of*



- uncertainty. in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 2017.
65. Williamson, K., M.A. Kennan, G. Johanson, and J. Weckert. *Data sharing for the advancement of science: Overcoming barriers for citizen scientists*. Journal of the Association for Information Science and Technology, 2016. **67**(10): p. 2392-2403.
66. Barbosu, S. and J.S. Gans. *Storm crowds: Evidence from Zooniverse on crowd contribution design*. Research Policy, 2022. **51**(1): p. 104414.
67. Hennon, C.C., K.R. Knapp, C.J. Schreck III, S.E. Stevens, J.P. Kossin, P.W. Thorne, P.A. Hennon, M.C. Kruk, J. Rennie, and J.-M. Gad  a. *Cyclone center: can citizen scientists improve tropical cyclone intensity records?* Bulletin of the American Meteorological Society, 2015. **96**(4): p. 591-607.
68. Chapman, L., C. Bell, and S. Bell. *Can the crowdsourcing data paradigm take atmospheric science to a new level? A case study of the urban heat island of London quantified using Netatmo weather stations*. International Journal of Climatology, 2017. **37**(9): p. 3597-3605.
69. Liu, K., C. Yang, Z. Li, M. Sun, J. Li, and C. Xu. *Climate@ Home: Utilizing Citizen Science for Climate Studies*. in *AGU Fall Meeting Abstracts*. 2013.
70. Tinati, R., M. Van Kleek, E. Simperl, M. Luczak-R  sch, R. Simpson, and N. Shadbolt. *Designing for citizen data analysis: A cross-sectional case study of a multi-domain citizen science platform*. in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2015.
71. Park, E.G., G. Burr, V. Slonosky, R. Sieber, and L. Podolsky. *Data rescue archive weather (DRAW): Preserving the complexity of historical climate data*. Journal of Documentation, 2018.
72. Lin, Y.-W., J. Bates, and P. Goodale. *Co-observing the weather, co-predicting the climate: Human factors in building infrastructures for crowdsourced data*. Science and Technology Studies, 2016. **29**(3): p. 10-27.
73. Kiryakov, A., B. Popov, I. Terziev, D. Manov, and D. Ognyanoff. *Semantic annotation, indexing, and retrieval*. Journal of Web Semantics, 2004. **2**(1): p. 49-79.
74. Mac  rio, C.G.N. and C.B. Medeiros. *Specification of a framework for semantic annotation of geospatial data on the web*. SIGSPATIAL Special, 2009. **1**(1): p. 27-32.
75. Hughes, A.J., J.D. Mornin, S.K. Biswas, L.E. Beck, D.P. Bauer, A. Raj, S. Bianco, and Z.J. Gartner. *Quanti. us: a tool for rapid, flexible, crowd-based annotation of images*. Nature methods, 2018. **15**(8): p. 587-590.
76. Gonzalez, A.L., D. Izidoro, R. Willrich, and C.A. Santos. *OurMap: Representing crowdsourced annotations on geospatial coordinates as Linked Open Data*. in *International Conference on Collaboration and Technology*. 2013. Springer.
77. Kaufmann, A., J. Peters-Anders, S. Yurtsever, and L. Petronzio. *Automated Semantic Validation of Crowdsourced Local Information – The Case of the Web Application "Climate Twins"*. 2013. Berlin, Heidelberg: Springer Berlin Heidelberg.
78. Santos, H. and V. Furtado. *A Service-oriented architecture for assisting the authoring of semantic crowd maps*. in *Brazilian Symposium on Artificial Intelligence*. 2012. Springer.
79. Heath, T. and C. Bizer. *Linked data: Evolving the web into a global data space*. Synthesis lectures on the semantic web: theory and technology, 2011. **1**(1): p. 1-136.
80. Goy, A., D. Magro, G. Petrone, M. Rovera, and M. Segnan. *Supporting Semantic Annotation in Collaborative Workspaces with Knowledge Based on Linked Open Data*. in *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*. 2016. Springer.
81. Debattista, J., C. Lange, S. Auer, and D. Cortis. *Evaluating the quality of the LOD cloud: An empirical investigation*. Semantic Web, 2018. **9**(6): p. 859-901.
82. Callaghan, W., J. Goh, M. Mohareb, A. Lim, and E. Law. *Mechanicalheart: A human-machine framework for the classification of phonocardiograms*. Proceedings of the ACM on Human-Computer Interaction, 2018. **2**(CSCW): p. 1-17.
83. Chan, J., J.C. Chang, T. Hope, D. Shahaf, and A. Kittur. *Solvent: A mixed initiative system for finding analogies between research papers*. Proceedings of the ACM on Human-Computer Interaction, 2018. **2**(CSCW): p. 1-21.
84. Li, T., K. Luther, and C. North. *Crowdia: Solving mysteries with crowdsourced sensemaking*. Proceedings of the ACM on Human-Computer Interaction, 2018. **2**(CSCW): p. 1-29.
85. Kensing, F. and J. Blomberg. *Participatory Design: Issues and Concerns*. Computer Supported Cooperative Work (CSCW), 1998. **7**(3): p. 167-185.
86. Schmidt, K. and L. Bannon. *Constructing CSCW: The first quarter century*. Computer supported cooperative work (CSCW), 2013. **22**(4): p. 345-372.
87. Dee, D.P., S.M. Uppala, A.J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, and d.P. Bauer. *The ERA-Interim reanalysis: Configuration and performance of the data assimilation system*. Quarterly Journal of the royal meteorological society, 2011. **137**(656): p. 553-597.
88. Tummolini, L. and C. Castelfranchi. *Trace signals: The meanings of stigmergy*. in *International Workshop on Environments for Multi-Agent Systems*. 2006. Springer.
89. Van den Besselaar, E.J., M. Haylock, G. Van der Schrier, and A. Klein Tank. *A European daily high-resolution observational gridded data set of sea level pressure*. Journal of Geophysical Research: Atmospheres, 2011. **116**(D11).
90. Liberato, M.L. *The 19 January 2013 windstorm over the North Atlantic: large-scale dynamics and impacts on Iberia*. Weather and Climate Extremes, 2014. **5**: p. 16-28.
91. Liberato, M.L., J.G. Pinto, I.F. Trigo, and R.M. Trigo. *Klaus–an exceptional winter storm over northern Iberia and southern France*. Weather, 2011. **66**(12): p. 330-334.
92. Daniel, F., P. Kucherbaev, C. Cappiello, B. Benatallah, and M. Allahbakhsh. *Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions*. ACM Comput. Surv., 2018. **51**(1): p. Article 7.
93. Oosterman, J., A. Nottamkandath, C. Dijkshoorn, A. Bozzon, G.-J. Houben, and L. Aroyo. *Crowdsourcing knowledge-intensive tasks in cultural heritage*. in *Proceedings of the 2014 ACM conference on Web science*. 2014.
94. Zhang, J., V.S. Sheng, J. Wu, and X. Wu. *Multi-class ground truth inference in crowdsourcing with clustering*. IEEE Transactions on Knowledge and Data Engineering, 2015. **28**(4): p. 1080-1085.
95. Schubert, E., J. Sander, M. Ester, H.P. Kriegel, and X. Xu. *DBSCAN revisited, revisited: why and how you should (still) use DBSCAN*. ACM Transactions on Database Systems (TODS), 2017. **42**(3): p. 1-21.
96. Goutte, C. and E. Gaussier. *A probabilistic interpretation of precision, recall and F-score, with implication for evaluation*. in *European conference on information retrieval*. 2005. Springer.
97. Kazai, G. and I. Zitouni. *Quality Management in Crowdsourcing using Gold Judges Behavior*, in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. 2016, Association for Computing Machinery: San Francisco, California, USA. p. 267–276.
98. Nielsen, J. *Usability engineering*. Boston: AP Professional, 1993.
99. Faulkner, L. *Beyond the five-user assumption: Benefits of increased sample sizes in usability testing*. Behavior Research Methods, Instruments, & Computers, 2003. **35**(3): p. 379-383.
100. Borsci, S., R.D. Macredie, J. Barnett, J. Martin, J. Kuljis, and T. Young. *Reviewing and Extending the Five-User Assumption: A Grounded Procedure for Interaction Evaluation*. ACM Trans. Comput.-Hum. Interact., 2013. **20**(5): p. Article 29.
101. Zhou, X., J. Tang, Y.C. Zhao, and T. Wang. *Effects of feedback design and dispositional goal orientations on volunteer performance in citizen science projects*. Computers in Human Behavior, 2020. **107**: p. 106266.

102. Rogstadius, J., V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic. *An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets*. in *Proceedings of the international AAAI conference on web and social media*. 2011.
103. Aitamurto, T. and J. Saldivar. *Motivating participation in crowdsourced policymaking: The interplay of epistemic and*

*interactive aspects*. *Proceedings of the ACM on Human-Computer Interaction*, 2017. **1**(CSCW): p. 1-22.

104. Sholler, D., I. Steinmacher, D. Ford, M. Averick, M. Hoyer, and G. Wilson. *Ten simple rules for helping newcomers become contributors to open projects*. *PLoS computational biology*, 2019. **15**(9): p. e1007296.

simulations for future climate scenarios. Additionally she is also interested on the impact of climate variability and extremes on society.

Additionally, she has participated as member of the research team of several FCT and EU funded projects such as ENAC, STORMEx and QSECA and has been a Member of the Management Committee of the COST ESSEM Action ES0604. She has also participated as PI in several hand-on projects (Concurso Ciência Viva VI) as well as in several tasks of the FP7 ACCENT (Atmospheric Composition Change) NETWORK OF EXCELLENCE from 2006-2009 and of the ACCENT Plus NETWORK OF EXCELLENCE since 2010. She has fostered several International collaborations which have resulted on several peer-reviewed publication

**Adriana S. Vivacqua** holds a degree from the Pontifical Catholic University of Rio de Janeiro (1993), a Master's degree in Computing from the Fluminense Federal University (1997), a Master's degree in Media Arts and Sciences from the Massachusetts Institute of Technology (1999) and a Ph. of Rio de Janeiro (2007) in co-supervision with the Université de Technologie de Compiègne, and post-doctoral by the Universidad Politécnica de Valencia. She is currently a professor at the Department of Computer Science at the Federal University of Rio de Janeiro, and works in the Graduate Program in Informatics, of which she is vice-coordinator.

**Andrea Grover** holds a PhD in Information Science and Technology from the Syracuse University School of Information Studies. Currently, she is an Associate Professor in Information Systems & Quantitative Analysis at the University of Nebraska at Omaha's College of Information Science & Technology. She studies public participation in data-intensive scientific collaboration (large-scale citizen science) and issues related to data management and technologies.

**Jeffrey P. Bigham** is an Associate Professor in the Human-Computer Interaction and Language Technologies Institutes in the School of Computer Science at Carnegie Mellon University. He also leads a Human-Centered Machine Intelligence Group at Apple, which works on research and applied projects in Accessibility, AI Fairness, ML Design, Learning Sciences, InfoVis, and Computational Understanding of UIs. He has received his B.S.E degree in Computer Science from Princeton University in 2003, and received his Ph.D. in Computer Science and Engineering from the University of Washington in 2009. He has received the Alfred P. Sloan Foundation Fellowship, the MIT Technology Review Top 35 Innovators Under 35 Award, and the National Science Foundation CAREER Award.

**Hugo Paredes** received the B.Eng. and Ph.D. degrees in computer science from the University of Minho, Braga, Portugal, in 2000 and 2008, respectively, and the Habilitation title from the University of Trás-os-Montes e Alto Douro (UTAD), Vila Real, Portugal, in 2016. Since 2003, he has been with UTAD, where he is currently an Associate Professor with Habilitation. In 2017, he was a Visiting Faculty with Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA. He is currently the Director of the Master Program with Informatics Engineering, UTAD, and a Member of the Managing Board of Ph.D. in Computer Science with UTAD. He is also an Assistant Coordinator with the Centre for Computer Graphics and Information Systems (CSIG), Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Porto, Portugal, where he is a Senior Researcher.

**Dennis Paulino** received the Master's Degree in Informatic Engineer, at University of Trás-os-Montes e Alto Douro (UTAD), Vila Real, Portugal in December 2018. He participated since September 2016 until December 2018 in the project NanoSTIMA RL2 - Passus Mobile, responsible for developing a system that makes exercise supervision of people with peripheral arterial disease. From January 2019 until November 2019, he participated in the project eCSAAP, responsible for the construction of a informatic system to help in the visualization and detection of meteorological phenomena. In December 2019 he ingress in PhD in Informatics at UTAD, with a scholarship financed by FCT.

**António Correia** is a former Microsoft Research fellow and works as a Research Assistant at INESC TEC, Porto, Portugal. He holds a Ph.D. degree in Computer Science from the University of Trás-os-Montes e Alto Douro (UTAD), Vila Real, Portugal. Furthermore, he formerly worked as a Visiting Postgraduate Researcher at University of Kent, Medway, UK. António holds more than ten years of experience in research and scientific writing, and his research interests are mainly in the fields of Human-Artificial Intelligence (AI) Interaction, Computer-Supported Cooperative Work (CSCW), and Information Systems (IS), and Science and Technology Studies (STS). He has authored or co-authored more than 50 publications, including journal articles, conference papers, book chapters, and posters. Moreover, he has also participated in research projects conducted at national and international level and has been executing functions as external reviewer and scientific committee member for top-tier journals and conferences covering aspects of computer science.

**Marcela Yagui** received the Master's Degree in Informatics at Fed. Univ. of Rio de Janeiro, Brazil, in April 2019. She has pursued the PhD degree in Informatics at Fed. Univ. of Rio de Janeiro.

**João Barroso** earned a doctorate in University of Trás-os-Montes e Alto Douro (UTAD), Vila Real, Portugal in 2002 in Electrical Engineering and held in 2008 the Habilitation in Informatics/Accessibility and became an Associate Professor in December 2012. He was Pro-Rector for Innovation and Information Management at UTAD from July 2010 to July 2013, Pro-Rector for Innovation and Technology Transfer from May 2017 and June 2020, and Vice-Rector for Innovation, Technology Transfer and Digital University since May 2021. He produced over 150 scientific papers, including book chapters, journal articles and articles in proceedings of scientific events. He supervised 40 postgraduate students (masters and doctorates). He was member of the research team in 35 research and development projects. He was member of several organizing committees of the international scientific meetings. In 2006 he directed the team that created the conference "Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (www.dsai.ws/2016) and in 2016 the conference Technology and Innovation in Sports, Health and Wellbeing (www.tishw.ws/2016). and Human Computer Interaction.

**Margarida Liberato** is an Assistant Professor at University of Trás-os-Montes e Alto Douro (UTAD and Researcher at the Instituto Dom Luiz (IDL), Faculty of Sciences, University of Lisbon. She has received her PhD in 2008 with an analysis of the extratropical stratosphere-troposphere circulation coupling using a 3D normal mode approach. Recently she has focused her research interests on the study of extratropical cyclones and storm-tracks variability, namely on natural hazards, extreme cyclones and cyclones with extreme impacts. She is the PI of STORMEx project: Mid-latitude North Atlantic Extreme Storms Variability: Diagnosis, Modeling Dynamical Processes and Related Impacts on Iberia. She is also participating in the international initiative of intercomparing extratropical cyclone detection and tracking algorithms (IMILAST - Intercomparison of mid latitude storm diagnostics) in order to assessing method-related uncertainties both using recent past and present climate reanalysis datasets and GCM