Experience from large-scale crowdsourcing via weather apps

Harald Kempf[1]

[1] German National Meteorological Service, Offenbach, Germany.

Author correspondence:
Dr. Harald Kempf,
Deutscher Wetterdienst
- Zentrale -
Frankfurter Straße 135
63067 Offenbach am Main
Germany.
Email: harald.kempf@dwd.de

URL: http://trauma.massey.ac.nz/issues/2021-3/AJDTS_25_3_Kempf.pdf

Abstract

This practice update presents the experience of launching a large-scale crowdsourcing feature using categorized user reports through an established weather app in Germany. Starting from the motivation for using crowdsourcing, this paper covers all development stages of the campaign from design through to legal considerations to the final rollout of the feature and first data analysis. Of particular focus is parameter choice and the possibility for automatic plausibility checks. We found that the newly-designed crowdsourcing feature was widely embraced by app users, which led to a very high number of reports. Analysing a sample dataset of approximately 660,000 observations from July to November 2020, we provide insight on data composition and quality of the reports as well as examples of the data integration into operational procedures. We offer some recommendations for potential new crowdsourcing campaigns based on our preliminary experience. Finally, we discuss possible future extensions as well as options to introduce standards and achieve an international data exchange.

Keywords: *Crowdsourcing, app, weather, best practice*

Crowdsourcing offers the chance to gather previously unavailable data on meteorological phenomena and thus greatly add to existing observation capabilities of meteorological services. Crowdsourcing as a form of citizen science, where members of the public are encouraged and supported to provide data, has the potential to mitigate problems and insufficiencies such as a lack of observation capacities (e.g., hail, snow depth) or sparse measuring networks. Furthermore, it can capture the actual impact on people of meteorological phenomena as a new type of measurement. This data offers the potential to connect local meteorological forecasts to local impact and thus greatly increase the usability and value of severe weather warnings.

Data obtained via crowdsourcing has an extremely wide range of potential applications. It can be employed to benefit forecasting and warning services, be used in assimilation and *nowcasting* (forecasting on a very short time scale), and as potential on-the-ground data for verification of forecasts and warnings. Consequently, a rising number of meteorological services launch new crowdsourcing campaigns, strengthen connections to voluntary weather observers and storm spotters, or make use of existing crowdsourced datasets. An overview of European meteorological services activities in this field is presented in Krennert et al. (2018) while organizations such as the European Meteorological Services Network (EUMETNET) and the World Meteorological Organization (WMO) are also developing inventories of existing crowdsourcing approaches to increase their visibility. Within the scope of this paper, we will focus on the aspect of crowdsourcing via categorized reports by untrained users with a focus on high-impact weather.

Design and Implementation

The German National Meteorological Service (DWD) operates an established weather app called WarnWetter, with approximately 10 million downloads and an active userbase of about one million users per month. This app was extended to include a new feature for crowdsourced weather reports by anonymous app users. While the basic version of the app is freely available on multiple app stores (e.g., https://play.google.com/store/apps/details?id=de.dwd.warnapp), the new feature could only be provided to users of the paid version of WarnWetter due to legal restrictions.

Designing the new crowdsourcing functionality required the consolidation of a wide array of requirements. Initially, stakeholder mapping was performed to identify the useful parameters to be obtained. These parameters of interest were investigated in regard to existing experience of other crowdsourcing actors (mostly other meteorological services) and possible existing standards for reporting (e.g., typical categories and thresholds). Ultimately, a selection of categories and values was made in a compromise between the demands of different

stakeholders (e.g., forecasters, model developers, special users) and a range of existing crowdsourcing approaches, in order to ensure the compatibility of potential future data exchanges.

Other important concerns were user friendliness and simplicity of the implementation. The overwhelming majority of users will most likely not be able to accurately report phenomena on a fine-grained meteorological scale. The final parameter set was partially composed of meteorological and impact-based parameters (see Table 1). User reports feature observations in standardized categories with corresponding values and special attributes. In addition, they can optionally report text comments and pictures of meteorological phenomena or impact.

Functionality and user interface design was implemented to allow for seamless integration into the existing app framework. The whole reporting process was required to be straightforward and fast in order to make it accessible for a wide range of potential users. Another major effort was the preparation of the legal framework around the crowdsourcing feature both in regard to collecting, storing, and processing potentially personal data and in regard to displaying raw user input, especially including user pictures, within a governmental app. Consequently, a strict opt-in is required to use the crowdsourcing feature. The according terms and conditions have to be accepted during registration or at a later point. Users can opt-out of the feature at any time.

Table 1
Overview of Parameter Categories as Presented in the App and Associated Plausibility Checks

Category	Value scale	Plausibility check
Lightning	4 levels, meteorological	Lightning or radar
Wind	5 levels, meteorological	Wind or radar data from numerical weather prediction (NWP)
Hail	6 levels, meteorological	Radar
Rain	5 levels, impact	Radar and cloud area fraction (CAF)
Slipperiness	3 levels, meteorological	NWP temperature
Snowfall	3 levels, meteorological	Radar or CAF and NWP temperature
Snowcover	5 levels, meteorological	NWP temperature
Cloudiness	4 levels, meteorological	CAF
Fog	3 levels, meteorological	-
Tornado	6 levels, impact	Radar

Note. In most cases numerical, meteorological values are used as a scale (e.g., time between strikes for lightning intensity). Wind initially had an impact-based scale, which was abandoned in favour of a meteorological scale (inspired by Beaufort) in order to better accommodate user reporting preferences.

To address potential privacy concerns, reporting was implemented quasi-anonymously. In order to prevent sabotage and harmful reports, a random device ID is associated with each report. Since this token is fully randomized and independent of personal data (such as other accounts or device hardware), it is not considered to be personalized information according to German law. It is also not possible to de-anonymize any users and observations are stored with only 250 metre spatial accuracy to avoid potential identification or tracking of users. Thus, overall the stored data does not qualify as “personal data“, which drastically simplifies the handling and offers full General Data Protection Regulation compliance.

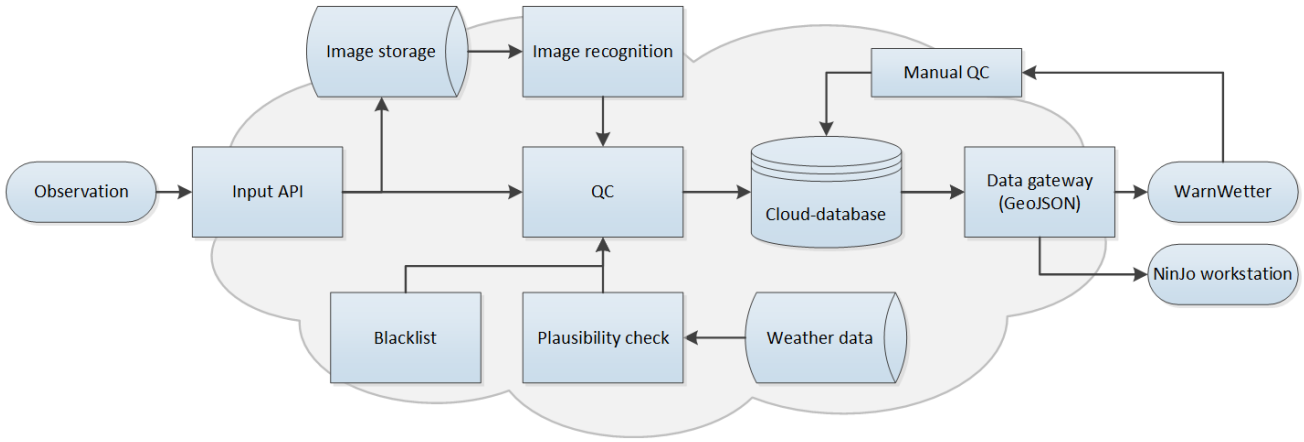
Users can optionally add pictures to their observations, submitted under a CC0-like licence which offers maximum flexibility to use and share the data. Due to peculiarities in German law, the CC0-licence could not be used directly and copyrights remain with the users. However, DWD gains all rights to use the data according to the terms and conditions.

Especially considering the potential display of illegal or harmful images in the app, further measures were taken in order to minimize this risk. Automatic unsafe content detection is applied to any user images. Images with clearly visible persons or body parts are flagged and not displayed in the app. Furthermore, reporting options for users have been implemented to instantly prevent any harmful images from being displayed.

To avoid potentially misleading false observations, a plausibility check was implemented in the application’s backend. The algorithm compares user observations to different datasets of existing meteorological observations and forecasts (predominantly radar measurements and NWP data) and automatically flags suspicious observations. Messages flagged as suspicious are not displayed to other users but are kept for further processing.

Data is stored in a cloud-hosted database and a web endpoint has been created which provides reports as GeoJSON (JavaScript Object Notation) files. Furthermore, an on-site data archive has been implemented at DWD. A schematic of the data processing is provided in Figure 1. At the end of the concept and development phase, extended testing of the new crowdsourcing feature was performed through pre-existing development channels.

Figure 1
Data Flow in the App Backend



Note. A plausibility check is applied to every observation in multiple steps. Most importantly, there is a comparison to existing weather data from radar, lightning measurements, satellite, and NWP. Observations are stored in a SQL-database and provisioned via a web interface in GeoJSON format.

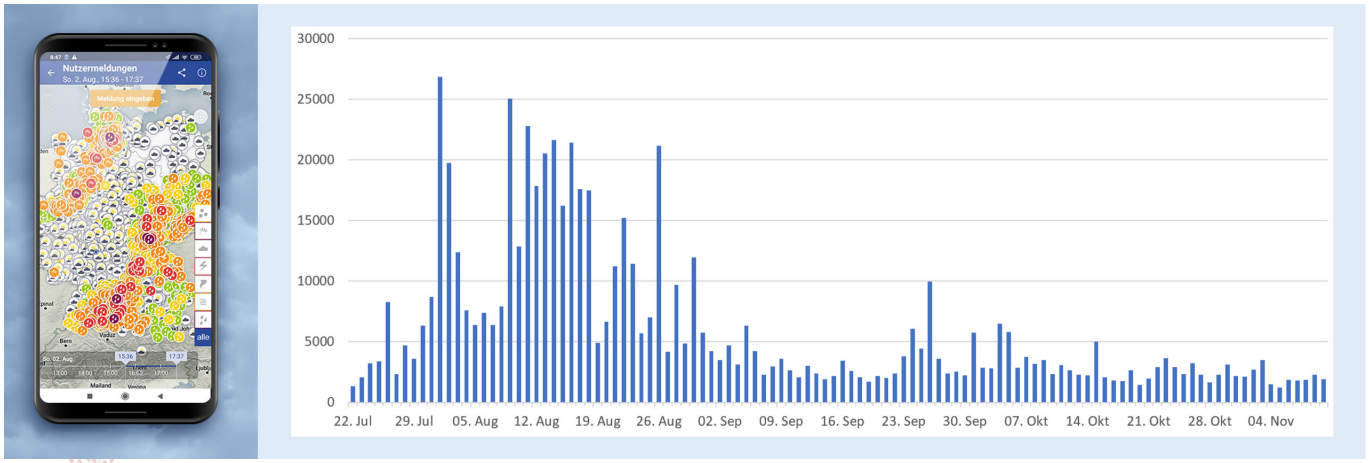
Rollout and Early Observations

The crowdsourcing feature was released to users using a staged rollout over the course of 1 week without any major technical difficulties. As the functionality was designed for intuitive usability, only a short introduction was provided to users in addition to minimal explanatory help text within the app.

Shortly after the full rollout, an overwhelming number of more than 26,000 messages per 24 hours was observed in a heavy rain event (as seen in Figure 2). Due to the very high number of messages and the maximum display period of 24 hours in the app, older smartphones were under serious stress when rendering all observations. As a quick response, the timeframe of messages to be displayed by default was limited to 1 hour in a point release. Further performance tweaks and new functionality were quickly provided in another full release. After the initial surge, the number of reports steadily decreased down to a baseline level of about 2,500 reports per 24 hours with expected spikes in severe weather situations (see Figure 2).

For a more detailed first analysis of observations, a subset recorded between the release of the feature on the 7th of July and the 11th of November 2020 was selected. This subset comprises about 660,000 observations from about 125,000 unique active contributors. Analysis revealed that the majority of observations were provided by casual (rather than consistent) users, with about 41% of users reporting only once. If this is due to users only testing out the new functionality or due to reporting only in a severe weather event is still to be evaluated. Another

Figure 2
Crowdsourcing Screen in the App WarnWetter and Number of Reports During a Heavy Rain Event on 2nd August 2020.



Note. Left side of figure: Crowdsourcing screen in the app WarnWetter as seen by users during a heavy rain event on 2nd of August 2020. Right side of figure: Total number of reports per day for the sample period from the official launch on 7th of July until the 11th of November.

47% of users reported up to 10 observations and about 7% up to 20. Of the remainder, 5% reported more than 20 times and about 0.5% of users contributed more than 100 reports each. A few users even actively scripted reports to be provided by their personal weather stations and webcams even though no API was provided.

About 8.5% of messages in the sample set included an accompanying image. The majority of images were reported in association with observations of cloudiness (about 80% overall). Nevertheless, a wide range of high impact situations featured in the user pictures (see Figure 3). User pictures were overall useful, especially for high impact situations such as slippery conditions. Only a few cases of false reports were observed (e.g., using images copied from the Internet) and almost no harmful reports (all of which were filtered by the unsafe content detection) even though reporting was de facto performed anonymously. Only 0.01% of images were reported by users to be problematic, and most of these reports were actually false positives.

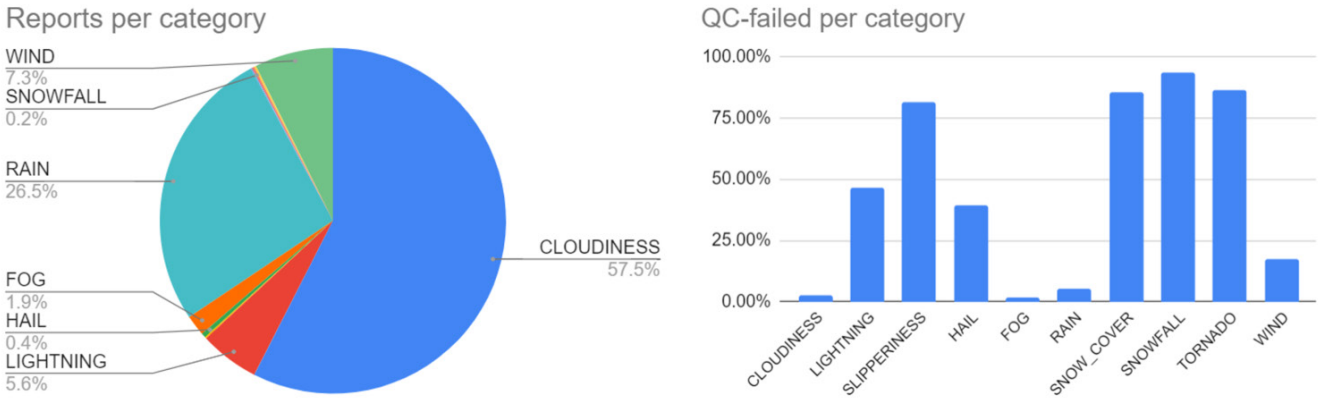
Meteorologically-false reports were flagged reasonably well by the automatic plausibility checks, due to the fact that many false reports were drastically wrong (e.g., reports of F3 tornadoes in calm weather). Only 0.4% of observations were reported at least once by other users to be not accurate, suggesting that the automatic control was sufficiently restrictive.

However, any plausibility checks need to be carefully crafted to allow for previously unknown data to be accepted when comparing to pre-existing conventionally measured or predicted data. As the sample period was mainly covering late summer and autumn, the observed high rejection rates for typical winter parameters such as snowfall, snow cover, and slipperiness are to be expected. For some categories such as lightning, hail, and wind however, the high number of flagged messages indicates that the initial choice of plausibility checks was too restrictive (see Figure 4). While this is not necessarily harmful (no false reports are displayed), the omission of potentially useful reports should be minimized.

Figure 3
Sample of User Pictures Provided Through the App

Note. Overall about 9% of messages included pictures, with a strong focus on cloudiness. Visual confirmation of the impact can be beneficial, especially for forecasters and users in civil defence.

Figure 4
Reports and Plausibility Check Failures per Category



Note. Left side of figure: Distribution of reports throughout the different categories. Right side of figure: Percentage of messages that failed the plausibility checks per category. Sample subset with 660,000 observations from July to November 2020.

Conclusions and Recommendations

Overall, the algorithm for automatic plausibility checks performed reasonably well. Manual plausibility checks could in principle be performed (e.g., by forecasters on duty). It would be beneficial to implement a two-stage process which combines an automatic flagging with a manual plausibility check. Manual inspection could thus be limited to suspicious reports only, making it much more feasible. Further automated plausibility checks via clustering would also be an option; however, the data is usually only available with sufficient density in urban regions. Automatic plausibility checks need to be carefully tuned and balanced for optimal performance between too permissive and too restrictive. In countries with strong seasonal differences, parameters for the checks might need to be split into independent summer and winter sets.

We also observed an interaction between reporting options offered to the users and plausibility checks. If citizens' willingness to report a meteorological phenomenon is high but there is no suitable reporting category provided, citizens may tend to misuse categories or thresholds. This is likely one reason behind the elevated level of wind observations flagged as suspicious (see Figure 4). Users were initially offered the option to report damaging effects of wind only, but they also wanted to report strong wind without damage. This led to a mismatch between observations and reports that was flagged by the plausibility check, as predicted wind speeds were not likely to cause any damage.

In response, the wind scale was adapted to match the user expectations more closely, moving away from an impact scale with three levels to a meteorological scale with five levels. A continuous monitoring of data quality and trends (e.g., high percentages of observations flagged by the automatic plausibility check) is strongly advised, especially in the early phases of a crowdsourcing campaign.

Any necessary changes in the reporting values or plausibility check parameters need to be carefully deliberated and meticulously tracked. Overall, the creation of a versioning system for these profiles seems advisable in order to keep track of all changes and to provide information on the exact profile used for a specific observation at any time. Especially for the use of crowdsourced observations in the context of numerical weather prediction and the operational production chain, the data and metadata quality are of extreme importance (Nipen et al., 2019).

When planning a new crowdsourcing effort, it is also necessary to reserve ample time for legal preparations during development, as challenges of data and privacy handling can be quite demanding depending on the local laws. Aiming for the minimal required amount of personal information and a *privacy by design* approach is often the key to being compliant to data protection laws, as illustrated throughout the current paper. Data minimization also has a positive effect on data handling and long-term storage.

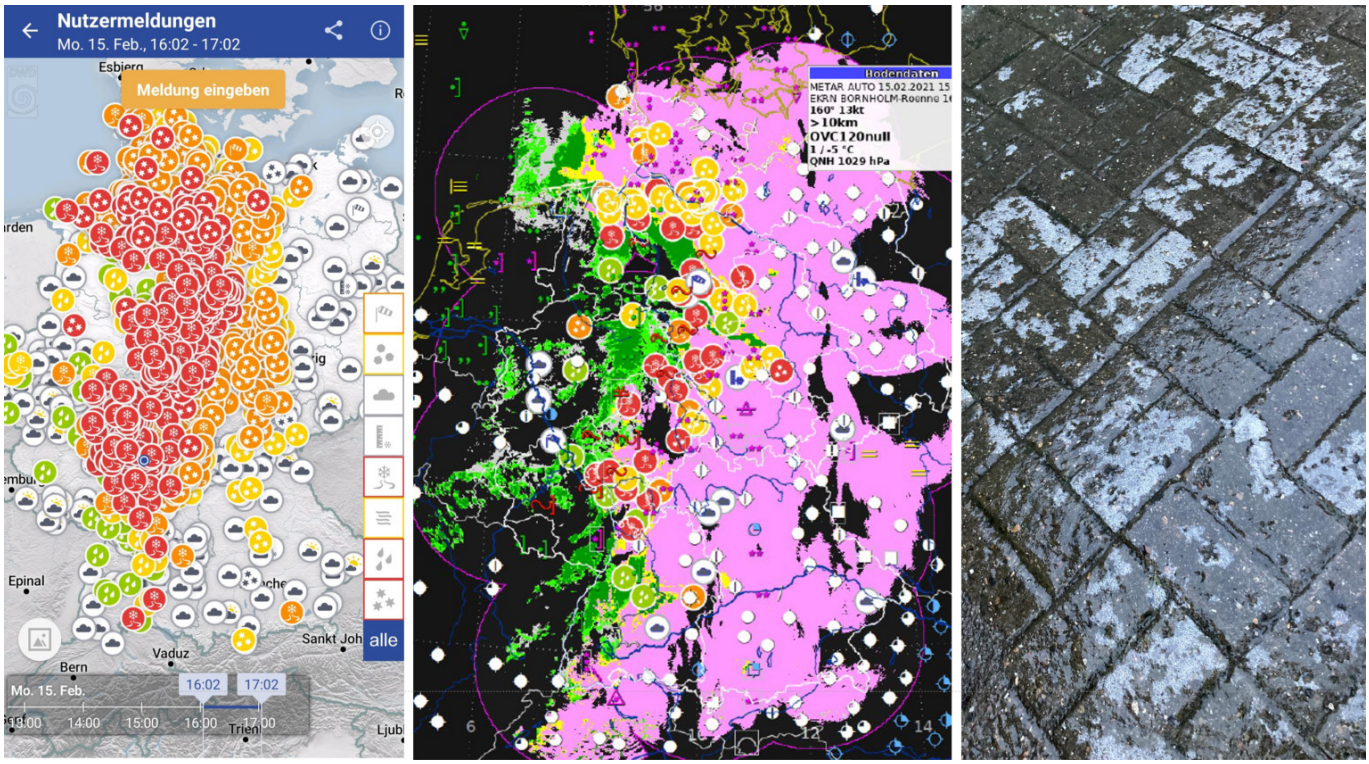
When launching a new crowdsourcing campaign, it is important to estimate the initial number of observations that will be sent in, especially since this amount will also strongly depend on the severity of the current weather. A scalable implementation of all required components is therefore paramount to provide sufficient capacity reserves and a satisfactory user experience.

Further, any new feature that is to be released for the use of the general public should have early large-scale testing followed by a small-scale rollout in order to avoid potential problems. Early testing by a dedicated user group also offers the chance for an overall more participatory nature of user involvement, potentially even actively including users in development cycles in a citizen science approach (Sturm & Martin, 2019). This approach is especially useful in order to find a good match for the offered reporting options between user expectations and expert needs. Key stakeholders such as emergency managers can be involved at this stage in order to tailor the functionality and results to their needs.

Close involvement can also have an educational aspect by increasing the sensibility of users to high-impact weather situations. Citizens can act as weather/impact observers via active queries ("Is there fog at your location?") or to verify forecast and warning accuracy ("Was there a thunderstorm at your location?"; "Was this warning accurate for you?"). Such participatory approaches might also offer better verification options, as the direct use of impact data in verification remains largely challenging due to a number of factors such as missing correct negatives (Crocker, 2018).

Another option for strengthening the involvement of users is aligned education programmes or gamification efforts. This can help to further increase the understanding of meteorological phenomena and severe weather risks and motivate users to maintain their reporting. Approaching special user groups such as trusted spotters, storm chasers, or citizens in civil defence can

Figure 5
Use of Crowdsourcing Data in Forecasting



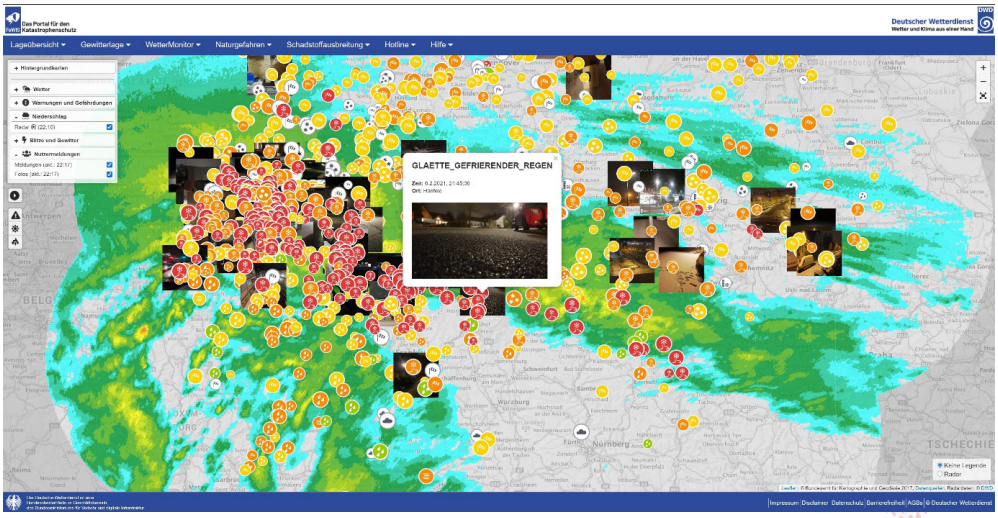
Note. Left panel: Situation during a freezing rain event in February 2021 as seen in the app. Middle panel: The NinJo forecaster workstation as a filtered dataset in conjunction with data on the precipitation phase. Right panel: Sample of user-provided impact images during the event.

offer potentially better observations as well as create a group of dedicated, trustworthy observers.

Data integration into existing systems and availability as datasets in common formats should be a high priority in order to make the best use of the data. Integration into operational systems also has the potential to provide an easy mechanism for manual quality control insofar as the systems can be extended to include according editing tools. Of central importance is the early integration into forecaster workstations, so that the data can be actively used to improve forecasts and warnings in high-impact situations. An example of this integration can be seen in Figure 5 for a high-impact freezing rain event. Both the general public and the forecasters benefited from the highly localized impact information gathered through crowdsourcing.

Crowdsourcing data was also directly provided to situation rooms and special users in civil defence via the fire brigades weather information (FeWIS) system, thus raising situational awareness and enabling a swifter and more precise response to the high impact event (see Figure 6). Especially for users in civil defence and emergency management, real-time impact information

Figure 6
Integration of Crowdsourcing Data Directly Within the FeWIS System for Special Users in Civil Defence



Note. Localized impact information can provide valuable insight into the current situation and the expected development during a high impact event (in this example, freezing rain and snowdrift).

is a key requirement, which in many cases cannot be provided by conventional meteorological measurements.

When displaying impact data from crowdsourcing, the choice of the right colour scale for visualization is of great importance. In our campaign, report categories were mapped to DWD´s warning thresholds and thus made use of the official four-colour scheme used in warnings. Preliminary analysis suggests that untrained users will have a tendency to report systematically stronger impacts than expected. Wind reports were a prime example of this tendency with users reporting hurricane force winds even in normal storms, potentially due to a subjectively felt higher impact or due to the rarity of the event. Consequently, it might be advisable to update the mapping of parameter colours if this mismatch becomes too strong, or to choose an independent colour scheme.

Full documentation including versioning metadata and in an accessible format such as GeoJSON facilitates the use of crowdsourced data by other actors and especially in research and development. Potential first steps include comparisons to other conventional observation sources to create trust in the new data source. This also makes it possible to draw on existing experience, for example in the comparison of data to radar observations (Barras et al., 2019). Especially in urban regions, the density of crowd observations will be very high (Meier et al., 2017) and accordingly the data can be of great use in climatological modelling of urban heat islands and city planning (Venter et al., 2020). Extensive experience exists for automated crowdsourcing (e.g., through private weather stations) – associated cross references can in part also be helpful for quality control in non-automated crowdsourcing (Fenner et al., 2017). If user images are part of the crowdsourcing effort, sophisticated data analysis tools such as machine learning can be employed for automatic classification and to build up impact databases. Through aligned datasets, the impact classification can be improved even further, especially for stakeholders in emergency management.

Involvement in international efforts to create standards is advisable, as the same platforms can also offer information on common best practice in crowdsourcing. Aligned efforts include the WMO High-Impact Weather (HIWeather) Citizen Science program and the EUMET crowdsourcing working group. Cooperation will also foster the potential for standardization, joint quality control techniques, and international data exchange.

References

Barras, H., Hering, A., Martynov, A., Noti, P.-A., Germann, U., & Martius, O. (2019). Experiences with >50,000 crowdsourced hail reports in Switzerland. *Bulletin of the American Meteorological Society, 100*(8), 1429-1440. https://doir.org/10.1175/BAMS-D-18-0090.1

Crocker, R. (2018). Can ad-hoc citizen observations be used to verify weather warnings? *Meteorologische Zeitschrift, 27*(6), 455-465. https://doi.org/10.1127/metz/2018/0879

Fenner, D., Meier, F., Bechtel, B., Otto, M., & Scherer, D. (2017). Intra and inter 'local climate zone' variability of air temperature as observed by crowdsourced citizen weather stations in Berlin, Germany. *Meteorologische Zeitschrift, 26*(5), 525-547. https://doi.org/10.1127/metz/2017/0861

Krennert, T., Pistotnik, G., Kaltenberger, R., & Csekits, C. (2018). Crowdsourcing of weather observations at national meteorological and hydrological services in Europe. *Advances in Science and Research, 15*, 71-76. https://doi.org/10.5194/asr-15-71-2018

Meier, F., Fenner, D., Grassmann, T., Otto, M., & Scherer, D. (2017). Crowdsourcing air temperature from citizen weather stations for urban climate research. *Urban Climate, 19*, 170-191. 10.1016/j.uclim.2017.01.006.

Nipen, T., Seierstad, I., Lussana, C., Kristiansen, J., & Hov, Ø. (2019). Adopting citizen observations in operational weather prediction. *Bulletin of the American Meteorological Society, 101*(1), 43-57. https://doi.org/10.1175/BAMS-D-18-0237.1

Sturm, U., & Tscholl, M. (2019). The role of digital user feedback in a user-centred development process in citizen science. *Journal of Science Communication, 18*(1), Article 3, 1-19. https://doi.org/10.22323/2.18010203

Venter, Z., Brousse, O., Esau, I., & Meier, F. (2020). Hyperlocal mapping of urban air temperature using remote sensing and crowdsourced weather data. *Remote Sensing of Environment, 242,* 111791. https://doi.org/10.1016/j.rse.2020.111791