AGU100 ADVANCING EARTH AND SPACE SCIENCE

Earth and Space Science

TECHNICAL REPORTS: DATA

10.1029/2018EA000454

Aurorasaurus Database of Real-Time, Crowd-Sourced Aurora Data for Space Weather Research

B. C. Kosar[1,2,3], Elizabeth A. MacDonald[2,3], Nathan A. Case[4], and Matthew Heavner[5]

[1]Department of Physics, The Catholic University of America, Washington D.C., USA, [2]NASA Goddard Space Flight Center, Greenbelt, MD, USA, [3]New Mexico Consortium, Los Alamos, NM, USA, [4]Department of Physics, Lancaster University, Lancaster, UK, [5]Los Alamos National Laboratory, Los Alamos, NM, USA

Key Points:
- Newly available Aurorasaurus database offers quality-controlled, citizen science, and social media reports of aurora
- The breadth of scientific utility of Aurorasaurus data is demonstrated
- Crowd-sourced aurora data is provided for model validation efforts of the space weather research community

Supporting Information:
- Supporting Information S1

Correspondence to:
B. C. Kosar,
bkosar@my.fit.edu

Citation:
Kosar, B. C., MacDonald, E. A., Case, N. A., & Heavner, M. (2018). Aurorasaurus database of real-time, crowd-sourced aurora data for space weather research. *Earth and Space Science*, *5*, 970–980. https://doi.org/10.1029/2018EA000454

Received 24 AUG 2018
Accepted 15 NOV 2018
Accepted article online 20 NOV 2018
Published online 21 DEC 2018

©2018. The Authors.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Abstract This technical report documents the details of Aurorasaurus citizen science data for the period spanning 2015 and 2016 as well as its routine data filtering protocols. Aurorasaurus citizen science data is a collection of auroral sightings submitted to the project via its website or apps and mined from social media. It is a robust data set and particularly abundant during strong geomagnetic storms when auroral precipitation models have the highest uncertainty. These data are offered to the scientific community for use through an open-access database in its raw and scientific formats, each of which is described in detail in this technical report. Furthermore, by demonstrating its scientific utility, we aim to encourage its integration into auroral research.

1. Introduction

Knowing the accurate location of the auroral oval with the progression of a geomagnetic storm is important for auroral research. Auroral oval predictions are generally based on the incorporation of data collected by various space-based particle detectors or imagers into empirical models (Evans, 1987; Hardy et al., 1985, 1989; Newell et al., 2009, 2010, 2014), however, the extent of their real-time prediction accuracy is unclear. Generally, they do not take into account contributions from substorms (explosive energy release within Earth's magnetic field) that can cause the auroral oval to expand and contract significantly within a few minutes. The time scale of dynamic auroral processes is faster than current operational models can predict. Auroral oval images obtained by space- and ground-based instruments provide more morphological detail in comparison to empirical model predictions. These observations are limited by coverage and typically the data are not readily available in real time due to image processing time requirements.

Aurorasaurus (MacDonald et al., 2015) is an innovative citizen science project focused on two fundamental scientific objectives: (1) collect real-time, ground-based aurora data from citizen scientists whose personal devices act as a form of soft-sensor and (2) incorporate this new type of data into scientific investigations related to aurora. Such citizen science and crowdsourcing data are becoming more common and important within space science (Cushley & Noël, 2014; Frissell et al., 2014).

2. Overview of Aurorasaurus Data

Aurorasaurus data are composed of direct reports submitted to the project via its website (aurorasaurus.org) and iOS and Android apps and tweets that are mined from Twitter via keyword searching and geotagging (Case, MacDonald, McCloat, et al., 2016). Direct reports can either be positive or negative, corresponding to whether or not the observer saw the aurora. The project has been live since September 2014. During the period of 2015–2016, the database compiled a total of 9,519 raw observations. The distribution of direct reports is shown in Figure 1a. The gray frame corresponds to the total number of direct reports collected by the project in 2015 (bar filled with diagonal lines) and 2016 (bar filled with dots). The green and the red frames show the number of positive and negative direct reports, respectively, for each year. Figure 1b shows the distribution of tweets that are mined from the Twitter social media platform. Twitter offers public access to its Application Programming Interface (API) through which interested communities can interact with their data. The pink frame corresponds to the total number of aurora-related tweets scraped from the Twitter search API. About 15% of these tweets, shown by the purple frame, contained geographical information (or location) with them. The geolocated tweets were presented to the Aurorasaurus community to vote on. The Aurorasaurus

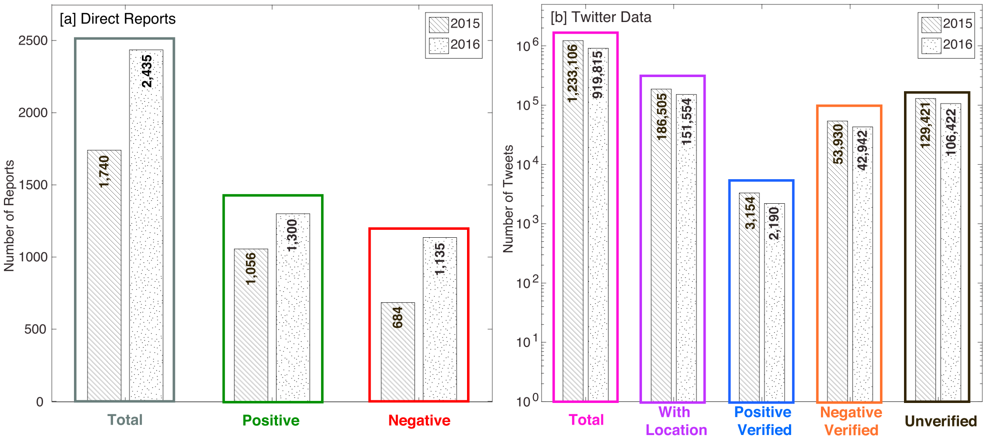

Figure 1. A distribution of 2015–2016 raw (a) direct reports collected via the project's website and apps and (b) data mined from the Twitter search Application Programming Interface.

project engages its community in tweet verification efforts by asking them to up or down vote the tweets presented on the Aurorasaurus platforms (website and apps). Tweets that are up-voted to be real-time auroral sightings are classified as *positive verified* tweets highlighted by the blue frame. The orange frame shows the number of *negatively verified* tweets indicating that they were not real-time auroral sightings or not actual auroral sightings at all and therefore, down-voted by the community (Case, MacDonald, McCloat, et al., 2016). The total number of negatively verified tweets for both years are significantly larger compared to positive verified tweets, reflecting the noise levels inherent in the Twitter data. The black frame shows that approximately 70% of the tweets were *unverified*. An earlier study by Case, MacDonald, Heavner, et al. (2015) showed that the number of reports submitted to Aurorasaurus scales with the strength of the geomagnetic activity. Even though 2015 was more active in terms of geomagnetic storms, the total number of reports submitted to the project increased by 40% in 2016. This demonstrates that the number of submissions is affected by other factors as well such as the growth of the size of the Aurorasaurus community, which grew from ~3,500 in 2015 to ~5,000 in 2016. A large number of the direct reports submitted during 2016 are negative which is expected and clearly emphasized by the 50% increase in number compared to 2015. The data mined from Twitter is consistently smaller in number during 2016 compared to 2015, likely due to declining geomagnetic activity. Even though the data scraped from the Twitter API are more numerous, only a small fraction of it is considered to be scientifically useful. Twitter is a unique source for robustly picking out relevant data during strong geomagnetic storm conditions (Case, MacDonald, Heavner, et al., 2015).

Aurorasaurus uses Postgres relational databases to store its data securely and organize it structurally (into rows and columns) for easy access via Structured Query Language query operations. Full database access is currently limited to project team members as well as the admin staff responsible for managing and maintaining it. Monthly data dumps from the database track data statistics and content. These files are stored at the New Mexico Consortium servers and are maintained by the technical staff of the institution. Recently, the Aurorasaurus database has increased its functionality by providing access to its data through an API for research and re-serving purposes. Before making this data set open access on Zenodo repository, interested research communities were granted limited access to Aurorasaurus data set upon request. Per our privacy policy, access to sensitive information such as the account details of the community members through the API is not permitted. Protecting the privacy of our community is a high priority of the project.

2.1. Description of the Content of Aurorasaurus Data Files

The hierarchical tree structure of the Aurorasaurus data files is shown in Figure 2. This data set is currently open access at Zenodo data repository (zenodo.org; Kosar, MacDonald, Case, & Heavner, 2018) along with 2017 data that is uploaded recently.

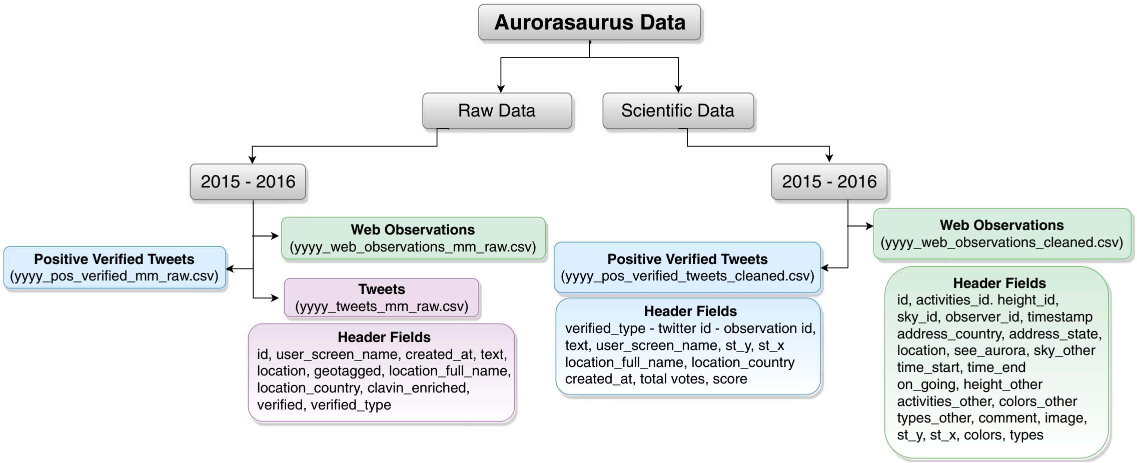
Aurorasaurus Data
Raw Data
Scientific Data
2015 - 2016
Web Observations
(yyyy_web_observations_mm_raw.csv)
Positive Verified Tweets
(yyyy_pos_verified_mm_raw.csv)
Tweets
(yyyy_tweets_mm_raw.csv)
Header Fields
id, user_screen_name, created_at, text, location, geotagged, location_full_name, location_country, clavin_enriched, verified, verified_type
2015 - 2016
Web Observations
(yyyy_web_observations_cleaned.csv)
Positive Verified Tweets
(yyyy_pos_verified_tweets_cleaned.csv)
Header Fields
verified_type - twitter id - observation id, text, user_screen_name, st_y, st_x location_full_name, location_country created_at, total votes, score
Header Fields
id, activities_id. height_id, sky_id, observer_id, timestamp address_country, address_state, location, see_aurora, sky_other time_start, time_end on_going, height_other activities_other, colors_other types_other, comment, image, st_y, st_x, colors, types

Figure 2. The hierarchical tree structure of the Aurorasaurus data files.

The two years (2015–2016) of shared data are either in their raw or scientific formats. Scientific data are the cleaned version of raw data by the processes described later in this section. For the raw data, three files are shared: Tweets (yyyy_tweets_mm_raw.csv or T-file), Positive Verified Tweets (yyyy_pos_verified_mm_raw.csv or PVT-file), and Web Observations (yyyy_web_observations_mm_raw.csv or WO-file). The *yyyy* and *mm* correspond to year and month of each year (i.e., 01 is January), respectively. WO-files contain reports submitted directly to the project via Aurorasaurus platforms. T-files contain all the aurora-related tweets that are mined from the Twitter search API via keyword searching such as *aurora* or *northern lights*. The Aurorasaurus server primarily filters this data by removing retweets, tweets containing spam terms, and Twitter users with aurora in their username. The content of the raw T- and PVT-files as well as cleaned PVT-files (yyyy_pos_verified_tweets_cleaned.csv) are described in Table 1.

Most of the data attributes found in T- and PVT-files are self-explanatory, however, it is worth giving a more detailed explanation of a few of them than what is given in Table 1. The allowed number of characters per tweet has traditionally been 140, as noted under *text* column, however, this has been updated to 280 characters per tweet starting late 2017. Therefore, Aurorasaurus data collected after 2017 will contain longer tweet texts. The location information (under *location* column) of the community member is saved as Well-Known Text format that is an alphanumeric representation of geometry on a map. This alphanumeric string can be converted to more readable geographic coordinates (latitude, st_y, and longitude, st_x) via query operations. If the location information is available, this means that the tweet has an embedded native geotag, therefore the geotagged column will be true ("t"). The geotagged tweet may also include location information in the textual format (e.g., Quincy, MA—United States) which is consecutively saved under location_full_name and location_country columns. In this scenario, the clavin_enriched column will show false ("f"). However, for tweets that do not come with a native geotag or a place name, we utilize an open source geoparsing software CLAVIN (Cartographic Location And Vicinity INdexer) (Greenbacker & Pinney, 2012-2014) to extract location information from the tweet text. In this scenario, the clavin_enriched column will be true (t).

PVT-files are subsets of T-files containing only the tweets that are positively verified as real-time aurora sightings by the members of the Aurorasaurus community. There are a total of 10 header fields in PVT-files and seven of them overlap with the content of T-files already described in Table 1. The four additional fields are st_y, st_x, total_votes, and score, two of which (st_y and st_x) are described earlier. Total_votes and score represent the number of votes cast on the tweet and the final score of the tweet (positive vote = +1 and negative vote = −1), respectively. The final score of a tweet must be greater than or equal to the threshold value set by the Aurorasaurus team to be classified as a positively verified tweet. Currently, this value is set to 2.

The Aurorasaurus project presents the citizen science community with a simple form to fill out for reporting their auroral sightings. The observer is asked to fill out the information on the location where the aurora was seen, and the observation period (start and end time of the observation). These geolocated and timestamped records of auroral visibility are frequently accompanied by optional, additional data describing the observed aurora and local environmental conditions (such as color, strength of the activity, location of the aurora in

Table 1
Description of Data Attributes Found in Raw T- and PVT-Files

Column header	Description	T or PVT?
id	Unique for each tweet	T, PVT
user_screen_name	Screen name of the community member who posted the tweet on Twitter	T, PVT
created_at	Posting time of the tweet	T, PVT
text	140 character text (frequently includes a link to the tweet window)	T, PVT
location	Well-known text (WKT) format community describing the location of the member	T
geotagged	Boolean (true or false) flag indicating if the tweets had a location embedded within them	T
location_full_name	Full location where the tweet was originated from	T, PVT
location_country	Country where the tweet was originated from	T, PVT
clavin_enriched	Boolean (true [t] or false [f]) flag indicating if CLAVIN software was used to extract the location information of the community member through the text of the tweet.	T
verified	Time when the tweet was verified	T
verified_type	If the tweet was verified, this field indicated the verification type (positive or negative)	T, PVT
st_y ($\pm$ 0–90°)	Latitude of the observation location	PVT
st_x ($\pm$0–180°)	Longitude of the observation location	PVT
total_votes	Number of votes cast on the tweet	PVT
score	Final score of the tweet (positive vote = +1 and negative vote = −1)	PVT

Note. Raw and cleaned version of PVT-file headers are identical to each other and they are a subset of column headers found in T-file with four additional fields. The distinction between T- and PVT-files is demonstrated in the last column. T = Tweets; PVT = Positive Verified Tweets.

the night sky, and auroral type). Raw WO-file have 24 data attributes that are identical to headers found in the cleaned version of this file (yyyy_web_observations_cleaned.csv) and they are described in Table 2. Web observations have the latitude and longitude information systematically obscured by a random amount of a kilometer or less, introducing an error of ±1 km, for privacy reasons.

The scientific data are the processed version of the raw data and maintain the same header fields. For ease of use, scientific data for all months for each year are combined into one file for positive verified tweets (yyyy_pos_verified_tweets_cleaned.csv) and one file for web observations (yyyy_web_observations_cleaned.csv).

Aurorasaurus, like any other citizen science project, exercises high data quality standards essential to the success of the project. Data are subject to thorough inspection for quality and integrity. Duplicate reports that are posted due to technical issues encountered during submission are filtered. Of interest to our primary scientific investigations are the negative reports with an indication of clear, unobscured view of the night sky. Therefore, negative reports that specify the sky condition to be *cloudy* or *bright* are removed from the data

Table 2
Description of Data Attributes in Raw and Cleaned WO-Files

Column header	Description
id	Unique for each observation
activities_id	Option for choosing the level of auroral activity (Quiet, Active, or Very Active)
height_id	Option for choosing the auroral height in the sky (Overhead, Northern Horizon, 45°N, 45°S, or Whole Sky)
sky_id (N/A for positive reports)	Option for choosing the sky condition during the observation (Cloudy, Clear, or Bright)
observer_id	Unique for each community member (blank for anonymous submissions)
timestamp (yyyy-mm-dd hh:mm:ss UT)	Observation submission time into Aurorasaurus platforms
address_country	Country of the observation
address_state	State of the observation (Effective for U.S. and Canada)
location	Well-known text (WKT) format describing the location of the community member
see_aurora	Boolean (true [t] or false [f]) flag indicating if the observer saw the aurora or not
sky_other	*Other* field allows observers to manually input description of the sky condition other than the options provided (see sky_id)
time_start (yyyy-mm-dd hh:mm UT)	Beginning time of the observation (15-min resolution)
time_end (yyyy-mm-dd hh:mm UT)	Ending time of the observation (15-min resolution)
on_going	Boolean (true [t] or false [f]) flag indicating if the auroral activity is continuing at the time of the report submission
height_other	*Other* field allows observers to manually input description of the auroral height in the sky other than the options provided (see height_id)
activities_other	*Other* field allows observers to manually input description of the level of auroral activity other than the options provided (see activities_id)
colors_other	*Other* field allows observers to manually input auroral colors observed other than the options provided (see colors_id)
types_other	*Other* field allows observers to manually input auroral types observed other than the options provided (see types)
comment	Allows observers to provide additional comments
image	If an auroral image captured by the observer was submitted to the server—yes [y] otherwise no [n]
st_y (±0−90°)	Latitude of the observation location (~1 km accuracy)
st_x (±0−180°)	Longitude of the observation location (~1 km accuracy)
colors	Option for choosing auroral colors (Red, Green, White, or Pink—community members can pick multiple colors)
types	Option for choosing auroral types (Discrete Arcs, Diffuse Glows, or Patches - community members can pick multiple types)

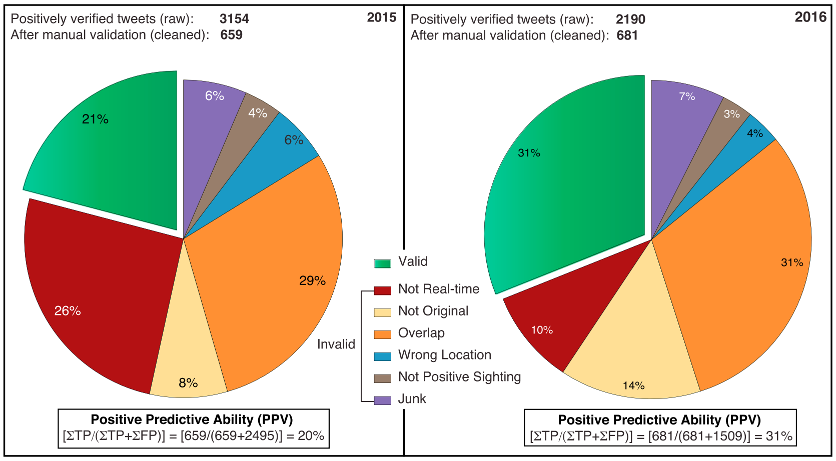

Figure 3. The distribution of positively verified tweets collected during 2015 and 2016.

set. However, negative reports that come with no indication of the sky condition (i.e., community member skips sky_id field) are counted as scientifically valuable data because the sky condition being clear is equally likely as being bright or cloudy.

Twitter data are also subject to rigorous processing for data quality by means of a three-step system: filtering, verification, and validation. As mentioned earlier, aurora-related tweets mined from Twitter are subject to filtering before being presented to the community on the Aurorasaurus platforms. Besides filtering, extracting meaningful signals from Twitter data requires verification and manual validation. Filtered tweets with location information are initially presented to the community members on Aurorasaurus platforms to verify if they are real-time aurora sightings. After exceeding a certain threshold (the final vote score should be greater than or equal to 2) a tweet is classified as a *positive verified tweet*. Verified tweets are checked annually following a predetermined set of rules to ensure their validity for detailed scientific analysis. The verification is a time consuming and labor intensive task that is primarily done by the Aurorasaurus team members and/or volunteers recruited under a standard protocol. Team members are the core group of scientists that are/were affiliated with the project. Volunteers are usually recruited from high school/undergraduate students through education and outreach activities of the project by the team members. Team members or volunteers involved in manual validation are required to read and understand the privacy policy of the project (http://aurorasaurus.org/privacy) prior to any sort of data handling or database access. Aurorasaurus community members are protected by our privacy policy. Personally identifiable information and data that requires proper crediting to their owner (such as images) are excluded from the public access.

The details of manual tweet verification are discussed in an earlier study (Case, MacDonald, McCloat, et al., 2016) based on the analysis of tweets collected during March and April 2015 that includes the period of St. Patrick's Day storm (Case, MacDonald, & Patel, 2015). The raw positively verified tweets are sifted through one at a time and they are divided into two major categories, valid or invalid. The valid category represents tweets that were identified correctly as real-time auroral sightings, while the invalid category is a collection of tweets that were misidentified as real-time auroral sightings by the Aurorasaurus community. The invalid category is further broken down into subcategories, that is, not real-time (red), not original (yellow), overlap (orange), wrong location (blue), not a positive sighting (gray), and junk (purple). The distribution of these categories for 2015 and 2016 data is shown in Figure 3. The description of each category can be found in the work of Case, MacDonald, McCloat, et al. (2016). True and false positives (TP and FP) refer to positively verified tweets that are valid and invalid, respectively. By utilizing the number of TP and FP, the positive predictive value for the tweet verification system was found to be 20% and 31% for 2015 and 2016, respectively. In other words, 20% and 31% of the tweets identified as positively verified in 2015 and 2016 were actually valid. There is an increase in this value for 2016; however, the source of this variance is not well understood. The increase is not attributable to sample size because although 2015 was more active (hence higher number of positively verified tweets) in comparison to 2016, the number of valid tweets is fewer.

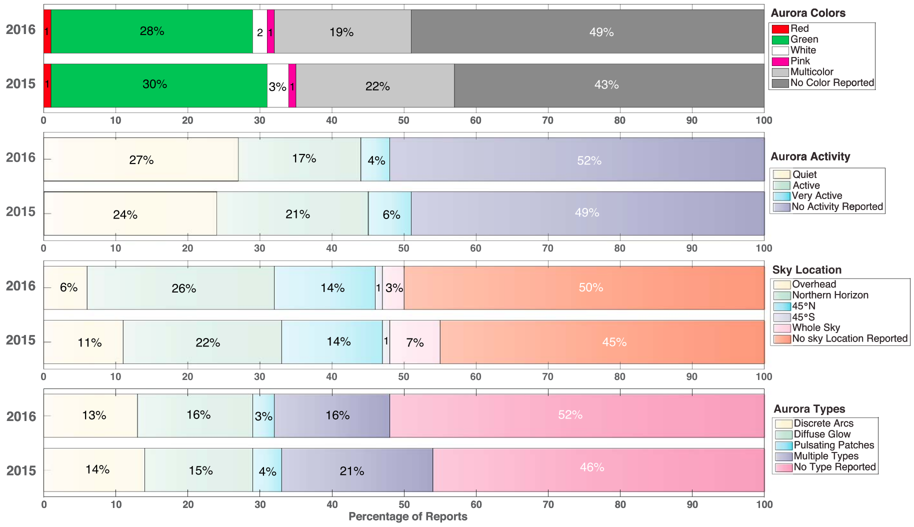

Figure 4. Description of the observed aurora by the citizen scientists during 2015–2016.

Figure 3 also shows that the percentage of the *not real-time* subcategory of invalid tweets is reduced in 2016. Identifying a tweet as real time or not requires detailed investigation of many aspects of that particular tweet. The procedure is a set of rules developed by the Aurorasaurus team members. For data quality assurance, team members and volunteers are trained on the same set of hundred tweets that were used during the project's first validation efforts. Because validating a large data set tends to be a time-consuming process, alternative techniques (such as machine learning algorithms) to speed up or eliminate manual validation efforts are being explored. The project currently has two years of data (2015–2016) validated for quality and readily available for scientific use. This data can be utilized for evaluation of existing models (Newell et al., 2009, 2014; Zhang & Paxton, 2008) and used as a new data source complementing the data-sparse field of Heliophysics.

2.2. Citizen Scientist Descriptions of Auroral Observations in 2015–2016

Of the 1,740 and 2,435 raw reports submitted in 2015 and 2016, 19.8% and 19.7% of them included an image of the observed aurora. Submitted auroral images are composed of smartphone photos of the back screen of a Digital Single-Lens Reflex camera, lower-quality smartphone images taken of the aurora directly, and high-quality postprocessed images. On average 52% of the reports also contain descriptive information about the observed aurora. If a community member skips one question on the form (e.g., color), they often skip the rest (i.e., type, sky location, activity). This is apparent in the percentages of each data attribute skipped being very similar. Figure 4 shows how citizen scientists described their observations during 2015–2016. Most of the observed aurora were either typical green auroral emission or multicolor (combination of green with other colors). The observed types are dominated by discrete arcs and diffuse glows or multiple types (combination of arcs, glows, and pulsating patches). Most observers described aurora being on the northern horizon or 45° above the horizon. The whole sky observations are sparse, which is likely due to the limited number of inhabitants at latitudes likely to see overhead aurora. Aurora was reported to be more active in 2015 (please see http://blog.aurorasaurus.org/?p=356) in comparison to 2016.

3. Scientific Utility of Aurorasaurus Database

The cleaned positive verified tweets and direct reports are subject to two more filters that are implemented in Interactive Data Language (IDL) codes. The plots shown in Figure 5 are produced for the time period of 01 January 2015 00:00:00 UT to 31 December 2016 23:59:59 UT. The first filter applied to the cleaned data files further checks to make sure the report times fall within this range. This filter removes only a few reports from the total (2 positive verified tweets and 12 positive reports). During submission, community members occasionally pick an incorrect time period (the difference between the end_time and the start_time) for their observations. The second filter removes positive/negative reports with an observation time period exceeding 3 hr, as they may contain an error or not be specific enough for analysis. In total, 214 positive and 18 negative reports are removed by filter two. Figures 5a and 5b are distributions of positive verified tweets and direct reports on a world map. These data are a collection of geolocated and timestamped signals of auroral visi-

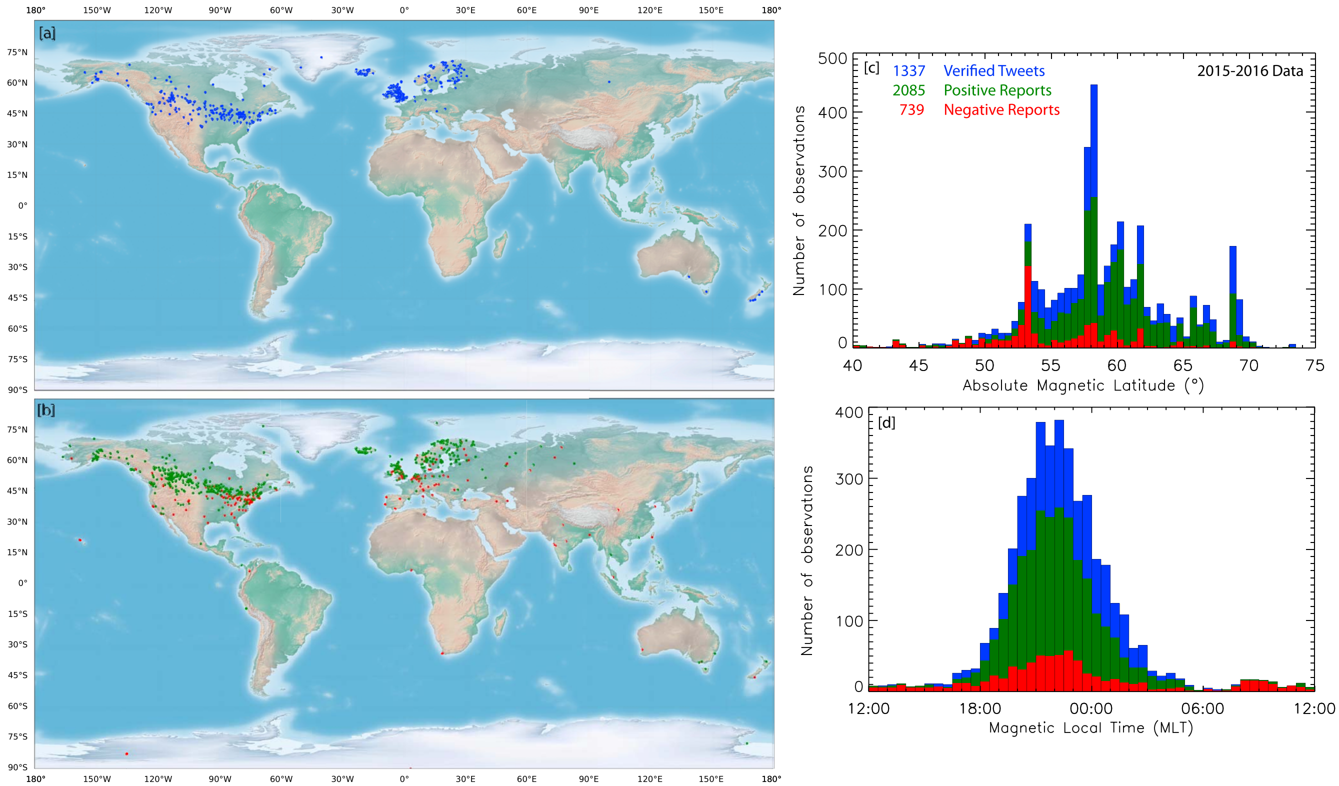

Figure 5. Distribution of validated (a) positive verified tweets and (b) web reports over the globe and the distribution of validated and filtered data as a function of (c) absolute magnetic latitude and (d) magnetic local time. Green and red filled circles correspond to positive and negative web reports, and blue filled circles correspond to positively verified tweets. The color code used for making the stacked bars refer to the same data types.

bility obtained from soft-sensors. These signals exhibit a sparse spatial organization with isolated regions of high signal density nested within low signal density distribution over the globe. Data coverage over land is reasonable, particularly around populated sectors of the high latitude regions of the northern hemisphere where aurora is typically visible. This scenario reverses to no data over the ocean and only a few points on the southern hemisphere due to the limited land area from which an aurora might be visible. With our systematic outreach efforts, particularly during strong geomagnetic activity, the Aurorasaurus community and contributed observations will continue to grow in the near future. In the world map shown in Figure 5b, there are a few data points (positive and negative reports) coming from very low latitude regions. While positive sightings at very low latitudes are highly unlikely, negative reports are still reasonable. Positive reports are most likely submitted by mistake or could be spam members submitting anonymously since there was no geomagnetic storm large enough to cause the auroral oval to expand that far south. This represents a minor caveat in positive reports.

Figures 5c and 5d show the distribution of Aurorasaurus reports submitted during 2015–2016, grouped by absolute magnetic latitude in 0.5° bins and magnetic local time in 30-min bins, respectively. The stacked green, red, and blue bars indicate the number of positive reports, negative reports, and verified tweets that fall into each bin. The distribution of this data as a function of absolute magnetic latitude indicates that the number of reports peak around ~58° latitude and span a wide range between 40° and 75° latitude. Aurorasaurus report submission hours span a range between 18:00 and 06:00 MLT with a peak around midnight. Most auroral models typically have the highest uncertainty during large geomagnetic storms when Aurorasaurus data are the most abundant. This unique data set can potentially help reduce this uncertainty.

3.1. Example Scientific Application

The scientific utility of this innovative and robust citizen science data collected by the Aurorasaurus project has been demonstrated in numerous publications across multiple disciplines. Case, MacDonald, Heavner, et al. (2015) is the first study showing the effectiveness of social media (Twitter) in detecting real-time auroral activity, specifically during strong geomagnetic disturbances. The large number of initial reports collected during the St. Patrick's Day storm of 2015 (Case, MacDonald, & Patel, 2015) by the Aurorasaurus platform were

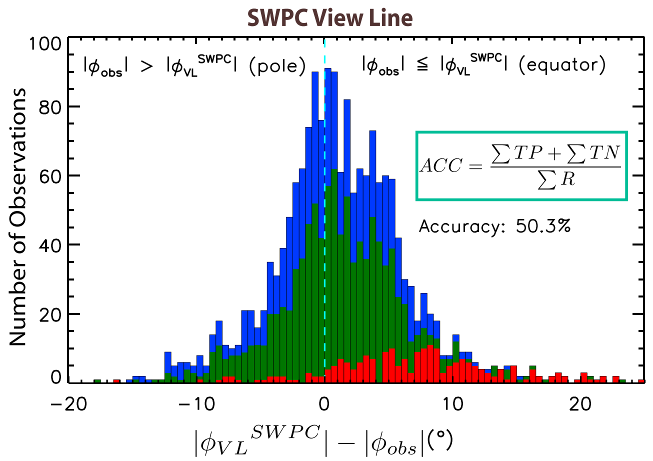

Figure 6. The differences in latitude between Aurorasaurus reports collected in 2015 and the SWPC view line at the same longitude are grouped into 0.5° bins. Stacked bars indicate number of each type of report falling into each interval. The color code used for the data types is the same as earlier. Approximately ~50% of the observations are reported from latitudes that are further equatorward of the view line estimated by the NOAA SWPC. The accuracy is calculated using true positive reports that include positively verified tweets (blue) and positive web reports (green) and true negative reports that include negative web reports (red). SWPC = Space Weather Prediction Center.

evaluated against the *view line*—an aurora forecast product of NOAA's Space Weather Prediction Center (SWPC) that is obtained using the predictions of Oval Variation, Assessment, Tracking, Intensity, and Online Nowcasting (OVATION) Prime 2010 auroral precipitation model and demonstrates the most southern latitude of the visible aurora. The results indicated that the latitudes of the majority of the citizen science reports were significantly equatorward of the view line latitudes predicted by the SWPC (Case, MacDonald, & Viereck, 2016). We note that the latitude of the citizen science reports solely represent the location of the observer submitting the report. The latitude is not derived using the location of the aurora in the sky. A recent case study (Kosar, MacDonald, Case, Zhang, et al., 2018) compared a subset of this data with the equatorial boundaries of the auroral oval at a fixed flux level obtained from the solar wind-driven OVATION Prime 2013 (OP-13) model (Newell et al., 2014) and the Kp-dependent Zhang-Paxton model (Zhang & Paxton, 2008). It was found that the OP-13 boundary is slightly more consistent with the citizen science data.

Global auroral particle precipitation is a result of coupling between the magnetosphere-ionosphere system that is driven by the external solar wind plasma input. The OVATION Prime 2013 (OP-13) auroral precipitation model uses a solar wind-magnetosphere coupling function to produce its high-resolution electron energy flux maps for the aurora. As described in Case, MacDonald, and Viereck (2016), this electron energy flux can be converted to a probability of visible aurora by scaling the summed precipitation energy flux (*j*) and adding an offset to it (i.e., P(A) = 10 + 8$\sum j$). In addition to this empirical conversion, NOAA's SWPC has a coarse estimate of a view line to account for the auroral height in the sky. The SWPC view line represents the lowest latitude where aurora should be visible. Aurorasaurus data are mostly clustered around the equatorial edge of the auroral oval hence offering useful data for assessing the accuracy of the view line. Following the earlier work (Case, MacDonald, & Viereck, 2016), outputs of the OVATION Prime 2013 model with a 15-min cadence were produced and the energy flux outputs were converted to percent probability of visible aurora. Figure 6 shows the distribution of Aurorasaurus data collected in 2015, grouped by latitude differences between Aurorasaurus data ($|\phi_{obs}|$) and SWPC view lines ($|\phi_{VL}^{SWPC}|$) into 0.5° bins. The accuracy is calculated using a statistical technique suggested by Machol et al. (2012), ACC = ($\sum$TP + $\sum$TN)/$\sum$R where $\sum$TP is the total number of true positive reports that fall within, $\sum$TN is the total number of true negative reports that fall outside of the view line, $\sum$R is the total number of reports. This equation yields an accuracy (ACC) of approximately 50.3% for the SWPC view line.

3.2. Aurorasaurus Database of Optical, Geotagged Auroral Imagery

Recent technological advancements have equipped citizen scientists with devices (smartphones, Digital Single-Lens Reflex cameras) that are capable of capturing high-quality image data. In the 2-year period of

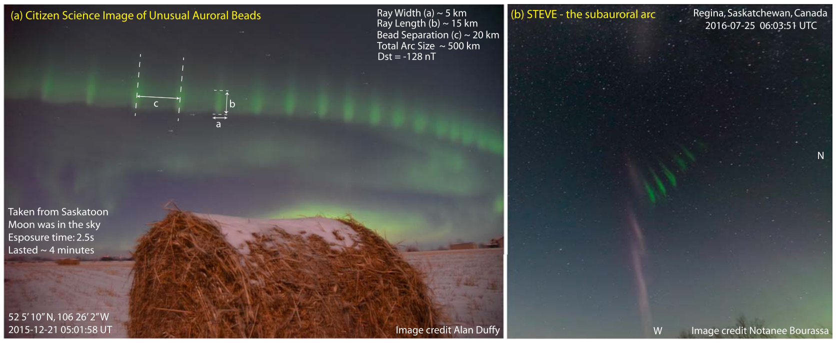

Figure 7. (a) Side view image of auroral beads observed during a geomagnetic storm from Saskatoon, Canada, using a Digital Single-Lens Reflex camera. The beads have a 20 km spacing based on star-tracking and analysis. (b) Image of STEVE and its accompanying green picket fence features forming south of the traditional auroral oval. STEVE = Strong Thermal Emission Velocity Enhancement.

2015–2016, a total of 823 auroral images have been submitted to the Aurorasaurus project accompanying the auroral sighting reports. We note that the image data are not shared on Zenodo due to the terms and conditions of the Aurorasaurus privacy policy. This database has permission for research use offering a unique collection of geotagged and optical auroral imagery as well as time lapse. Even though image sequences captured by the citizen scientists are rare, they are particularly useful in visualizing temporal and spatial dynamics of auroral arcs during geomagnetic storms. One example are auroral beads that are repeating patterns or structures within the auroral arcs. Typically, scientific instruments such as imagers on-board satellites or all-sky cameras capture them from above or below and may not have the resolution for fine scale structures. Citizen science images, such as the one shown in Figure 7a, provide scientists with a new set of data obtained from ground but from a different perspective and resolution. This particular side profile image of auroral beads allowed us to determine dimensions of an individual upright ray (width ~5 km and length ~15 km), the separation between two arbitrarily selected rays (~20 km), and the approximate total arc size within the field of view (~500 km) using star field analysis. The image sequence of this particular event allowed us to observe the direction of motion of individual rays. Citizen scientists collecting images of auroral arcs such as these provide new pieces of information about aurora that contribute to research interests of the space weather community. The Aurorasaurus blog has posted an article (http://blog.aurorasaurus.org/?p=398) on auroral beads featuring this particular image and discussing it relative to images of auroral beads captured by all-sky imagers and instruments on-board Earth-orbiting satellites (Henderson, 2008; Kalmoni et al., 2015).

A collaborative research opportunity between the Aurorasaurus citizen science network and auroral researchers has recently led to the discovery of an optical signature of a new subauroral phenomena (see Figure 7b) — Strong Thermal Emission Velocity Enhancement (Gallardo-Lacourt et al., 2018; MacDonald et al., 2018). This transient structure forms equatorward of the traditional auroral oval and displays a purplish color that is not typical of an auroral emission. In the declining period of solar maximum, these phenomena have been frequently caught on citizen scientists' cameras and submitted to the Aurorasaurus project. With an expanding Aurorasaurus community, this image database will continuously grow to allow opportunities for detailed analysis of Strong Thermal Emission Velocity Enhancement in the near future.

4. Conclusions

The Aurorasaurus project provides curated citizen science aurora data, particularly abundant during strong geomagnetic storms, as a useful resource for the space weather research community. Currently, 2 years (2015–2016) of data are available for scientific use due to data validation challenges. Alternative solutions for automating this effort is a work in progress and an important future step for the Aurorasaurus project. The newly emerging fields of artificial intelligence and machine learning offers algorithms (natural language processing, classification, etc.) that may be well-suited for the tweet validation efforts of the project.

To demonstrate the scientific utility of this data set, Aurorasaurus reports are compared with the OVATION-driven view line predictions of NOAA SWPC for 2015. Aurorasaurus reports are mostly clustered around the equatorial edge of the auroral oval, hence offering a useful data set for assessing accuracy. We find that ~50% of the observations are reported from the latitudes that are further equatorward of the view line estimated by NOAA SWPC. This unique data set has a great potential for validating, improving, and complementing existing models for auroral oval predictions and specifications. Emerging computational methods based on data-model integration offer new insights that could potentially improve real-time assessment and space weather prediction when citizen science data are combined with traditional sources. A future study will focus on developing a state-of-the-art auroral assimilative model that combines observational data (citizen science reports) with existing empirical models. Once developed, this assimilative model will provide feedback to model validation and ionospheric conductance challenges introduced by the NASA Community Coordinated Modeling Center (https://ccmc.gsfc.nasa.gov).

The Aurorasaurus database also offers high-quality images and time-lapse sequences of aurora captured by the community members. This geotagged image database contains a new set of data obtained from the ground but from a different perspective in comparison to ground- and space-based scientific equipment. This image database is a valuable complement to current scientific research and also provides opportunities for new discoveries advancing our understanding of the night sky.

Acknowledgments

This material is based upon work supported, in part, by the National Science Foundation (NSF) under grant 1344296 and NASA CAN award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF. N.A.C. was supported during this study by STFC grant ST/M001059/1. The OVATION Prime output and associated view line were kindly supplied by the Space Weather Prediction Center, Boulder, CO, National Oceanic and Atmospheric Administration (NOAA), U.S. Department of Commerce. The output can be freely downloaded from the NOAA SWPC product pages (http://www.swpc.noaa.gov/products/aurora-30-minute-forecast). Reid Priedhorsky provided the initial software engineering and contributions to the Aurorasaurus vision. Ideum and David Kingman coded the database, website, and apps described here. We would like to thank Sean McCloat, Mike Cook, and Madison Smith for their valuable help with tweet validation efforts of the Aurorasaurus project.

References

Case, N. A., MacDonald, E. A., Heavner, M., Tapia, A. H., & Lalone, N. (2015). Mapping auroral activity with Twitter. *Geophysical Research Letters*, *42*, 3668–3676. https://doi.org/10.1002/2015GL063709

Case, N. A., MacDonald, E. A., McCloat, S., Lalone, N., & Tapia, A. (2016). Determining the accuracy of crowdsourced tweet verification for auroral research. *Citizen Science: Theory and Practice*, 13.

Case, N., MacDonald, E., & Patel, K. (2015). Aurorasaurus and the St. Patrick's Day storm. *Astronomy and Geophysics*, *56*(3), 3–13.

Case, N. A., MacDonald, E. A., & Viereck, R. (2016). Using citizen science reports to define the equatorial extent of auroral visibility. *Space Weather*, *14*, 198–209. https://doi.org/10.1002/2015SW001320

Cushley, A. C., & Noël, J.-M. (2014). Ionospheric tomography using ADS-B signals. *Radio Science*, *49*, 549–563. https://doi.org/10.1002/2013RS005354

Evans, D. S. (1987). Global statistical patterns of auroral phenomena. In *Proceedings of Quantitative models of Magnetospheric-Ionospheric Coupling Processes* (pp. 325–330). Kyoto, Japan.

Frissell, N. A., Miller, E. S., Kaeppler, S. R., Ceglia, F., Pascoe, D., Sinanis, N., et al. (2014). Ionospheric sounding using real-time amateur radio reporting networks. *Space Weather*, *12*, 651–656. https://doi.org/10.1002/2014SW001132

Gallardo-Lacourt, B., Liang, J., Nishimura, Y., & Donovan, E. (2018). On the origin of steve: Particle precipitation or ionospheric skyglow? *Geophysical Research Letters*, *45*, 7968–7973. https://doi.org/10.1029/2018GL078509

Greenbacker, C., & Pinney, T. (2012-2014). Clavin [software]. Berico Technologies.

Hardy, D. A., Gussenhoven, M. S., & Brautigam, D. (1989). A statistical model of auroral ion precipitation. *Journal of Geophysical Research*, *94*(A1), 370–392.

Hardy, D. A, Gussenhoven, M. S., & Holeman, E. (1985). A statistical model of auroral electron precipitation. *Journal of Geophysical Research*, *90*(A5), 4229–4248.

Henderson, M. G. (2008). Observational evidence for an inside-out substorm onset scenario. *Annales Geophysicae*, *27*, 2129–2140.

Kalmoni, N. M., Rae, I. J., Watt, C. E., Murphy, K. R., Forsyth, C., & Owen, C. J. (2015). Statistical characterization of the growth and spatial scales of the substorm onset arc. *Journal of Geophysical Research: Space Physics*, *120*, 8503–8516. https://doi.org/10.1002/2015JA021470

Kosar, B. C., MacDonald, E. A., Case, N. A., & Heavner, M. (2018). Aurorasaurus Real-Time Citizen Science Aurora Data (Version v1.0) [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.1255196

Kosar, B. C., MacDonald, E. A., Case, N. A., Zhang, Y., Mitchell, E. J., & Viereck, R. (2018). A case study comparing citizen science aurora data with global auroral boundaries derived from satellite imagery and empirical models. *Journal of Atmospheric and Solar-Terrestrial Physics*, *177*, 274–282.

MacDonald, E. A., Case, N. A., Clayton, J. H., Hall, M. K., Heavner, M., Lalone, N., et al. (2015). Aurorasaurus: A citizen science platform for viewing and reporting the aurora. *Space Weather*, *13*, 548–559. https://doi.org/10.1002/2015SW001214

MacDonald, E. A., Donovan, E., Nishimura, Y., Case, N. A., Gillies, D. M., Gallardo-Lacourt, B., et al. (2018). New science in plain sight: Citizen scientists lead to the discovery of optical structure in the upper atmosphere. *Science Advances*, *4*(3), eaaq0030. https://doi.org/0.1126/sciadv.aaq0030

Machol, J. L., Green, J. C., Redmon, R. J., Viereck, R. A., & Newell, P. T. (2012). Evaluation of OVATION Prime as a forecast model for visible aurorae. *Space Weather*, *10*, S03005. https://doi.org/10.1029/2011SW000746

Newell, P. T., Liou, K., Zhang, Y., Sotirelis, T., Paxton, L. J., & Mitchell, E. J. (2014). OVATION Prime-2013: Extension of auroral precipitation model to higher disturbance levels. *Space Weather*, *12*, 368–379. https://doi.org/10.1002/2014SW001056

Newell, P. T., Sotirelis, T., & Wing, S. (2009). Diffuse, monoenergetic, and broadband aurora: The global precipitation budget. *Journal of Geophysical Research*, *114*, A09207. https://doi.org/10.1029/2009JA014326

Newell, P. T., Sotirelis, T., & Wing, S. (2010). Seasonal variations in diffuse, monoenergetic, and broadband aurora. *Journal of Geophysical Research*, *115*, A03216. https://doi.org/10.1029/2009JA014805

Zhang, Y., & Paxton, L. J. (2008). An empirical Kp-dependent global auroral model based on TIMED/GUVI FUV data. *Journal of Atmospheric and Solar-Terrestrial Physics*, *70*(8), 1231–1242.