

scientific data

OPEN
DATA DESCRIPTOR



Structured dataset of reported cloud seeding activities in the United States (2000–2025) using an LLM

Jared Joseph Donohue¹  & Kara D. Lamb²

Cloud seeding, a weather modification technique used to increase precipitation, has been practiced in the western United States since the 1940s. However, comprehensive datasets are not currently available to analyze these efforts. To address this gap, we present a structured dataset of reported cloud seeding activities in the U.S. from 2000–2025, including the project name, year, season, state, operator, seeding agent, apparatus used for deployment, stated purpose, target area, control area, start date, and end date. Combining our multi-stage PDF-to-text extraction pipeline with OpenAI's *o3* large language model (LLM), we processed 832 historical reports from the National Oceanic and Atmospheric Administration (NOAA). The resulting dataset demonstrates 98.38% estimated accuracy, based on manual review of 200 randomly sampled records, and is publicly available on Zenodo. This dataset addresses the gap in cloud seeding data and demonstrates the potential for LLMs to extract structured information from historical environmental documents. More broadly, this work provides a scalable framework for unlocking historical data from scanned documents across scientific domains.

Background & Summary

Cloud seeding is a weather modification technique used to enhance precipitation, typically in regions experiencing drought, water scarcity, or to support snowpack accumulation. In the United States, cloud seeding experiments began in the 1940s¹, and systematic reporting was introduced with the Weather Modification Reporting Act of 1972². The Act requires individuals and organizations to notify the U.S. Department of Commerce at least 10 days before and after conducting weather modification activities, using Form 17-4 (Initial Report on Weather Modification Activities) and Form 17-4A (Interim Activity Reports and Final Report). These forms are intended to be submitted to NOAA and archived on the NOAA Weather Modification website³. However, only reports from 2000 to 2025 are currently available online. Records from earlier years are not accessible through NOAA's website and have not yet been located elsewhere. As a result, our dataset covers only the 2000–2025 period, representing the full set of publicly accessible reports to date.

While the NOAA reports are made publicly available online, they are stored individually as scanned PDFs with inconsistent formatting and structure. This lack of standardization makes it difficult to extract key information at scale, or integrate the reports into other structured datasets. As a result, data on cloud seeding in the U.S., such as locations, dates, purposes, cloud seeding agent, and agent deployment methods, remains largely inaccessible for analysis.

To solve this data access problem, we created a structured dataset from 832 NOAA cloud seeding reports spanning from 2000 to 2025. Our work transforms these scattered, inconsistently formatted documents into a single, structured CSV file that enables the analysis of U.S. weather modification activities over time. The dataset achieves an estimated 98.38% accuracy across all extracted fields, based on a manual evaluation of 200 randomly sampled records (out of the total population of 832), each reviewed by two independent human annotators against the original PDFs.

This data extraction method can be extended to similar government-mandated environmental reporting systems, such as those related to water usage, air quality monitoring, or land management, many of which exist in PDF form with highly variable formatting. By transforming previously inaccessible records into structured data, this work lays the groundwork for evidence-based research and policy evaluation in an area of growing environmental interest⁴.

¹Data Science Institute, Columbia University, New York, NY, USA. ²Department of Earth and Environmental Engineering, Columbia University, New York, NY, USA. e-mail: jjd2203@columbia.edu

Complete in accordance with instructions on reverse and forward copy:

TO: National Oceanic and Atmospheric Administration
1315 East-West Highway, SSMC-2, Room 11554
Silver Spring, MD 20910

Form Approved: OMB No. 0648-0025 Expires 09/03/2004
NOAA FORM 17-4 U.S. DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
INITIAL REPORT ON WEATHER MODIFICATION
ACTIVITIES (PL. 205, 92nd CONGRESS)

1. PROJECT OR ACTIVITY DESIGNATION, IF ANY
Eastern San Juan Program

2. DATES OF PROJECT
a. DATE FIRST ACTUAL WEATHER MODIFICATION ACTIVITY IS TO BE UNDERTAKEN 11-01-2005

3. PURPOSE OF PROJECT OR ACTIVITY
Winter snowpack and precipitation augmentation

4. (a) SPONSOR
NAME: L. M. Hjermstad
See attached pages

5. (b) OPERATOR
NAME: PHONES NUMBER (970) 247-8813
AFFILIATION: Durango Consultants, LLC

6. TARGET AND CONTROL AREAS (See Instructions)

LOCATION	SIZE OF AREA	LOCATION	SIZE OF AREA
Eastern San Juan Mts	800 SQ.MI	Northern New Mexico	500 SQ.MI
		Mountains Along Continental Divide	

7. DESCRIPTION OF WEATHER MODIFICATION APPARATUS, MODIFICATION AGENTS AND THEIR DISPERSAL RATES, THE TECHNIQUES EMPLOYED, ETC. (See Instructions)

Apparatus: (12) Twelve manually controlled ground-based cloud nuclei generators.

Modification Material: Acetone solution of 5% silver iodide & 1.25% sodium iodide. Rate of dispensing material will range from 3 Gm / Hr to 25 Gm / Hr depending on cloud-base moisture & temperature, cloud-base height and cloud type.

Precipitation Measurements: Standard NWS precipitation observations along with SCS Snotel observations in both seeded and control areas.

NOAA Form 17-4

FILENAME: Eastern San Juan Program_07-1352_11.01.2006-03.31.2007.pdf
===== NOAA FORM 17-4: INITIAL REPORT ON WEATHER MODIFICATION ACTIVITIES =====
07-1352
Complete in accordance with instructions on reverse and forward copy and Approved: OMB No. 0648-0025 Expires 09/03/2004
TO: National Oceanic and Atmospheric Administration (4-81) NATL OCEANIC AND ATMOSPHERIC ADMIN.
Office of Oceanic and Atmospheric Research Silver Spring, MD 20910 SSMC-3 Room 11554 ACTIVITIES (PL. 205, 92nd CONGRESS)
INITIAL REPORT ON WEATHER MODIFICATION
1. PROJECT OR ACTIVITY DESIGNATION, IF ANY a. DATES OF PROJECT
Eastern San Juan Program b. DATE FIRST ACTUAL WEATHER MODIFICATION ACTIVITY IS TO BE UNDERTAKEN Nov. 1, 2006
3. PURPOSE OF PROJECT OR ACTIVITY b. EXPECTED TERMINATION DATE OF WEATHER

4. (a) SPONSOR
NAME: L. M. Hjermstad
See attached pages

5. (b) OPERATOR
NAME: PHONES NUMBER (970) 247-8813
AFFILIATION: Durango Consultants, LLC

6. TARGET AND CONTROL AREAS (See Instructions)

LOCATION	SIZE OF AREA	LOCATION	SIZE OF AREA
Eastern San Juan Mts	800 SQ.MI	Northern New Mexico	500 SQ.MI
		Mountains Along Continental Divide	

7. DESCRIPTION OF WEATHER MODIFICATION APPARATUS, MODIFICATION AGENTS AND THEIR DISPERSAL RATES, THE TECHNIQUES EMPLOYED, ETC. (See Instructions)

Apparatus: (12) Twelve manually controlled ground-based cloud nuclei generators.

Modification Material: Acetone solution of 5% silver iodide & 1.25% sodium iodide. Rate of dispensing material will range from 3 Gm / Hr to 25 Gm / Hr depending on cloud-base moisture & temperature, cloud-base height and cloud type.

Precipitation Measurements: Standard NWS precipitation observations along with SCS Snotel observations in both seeded and control areas.

extracted PDF text

A	B	C	D	E	F	G	H
1	filename	year	season	state	agent	apparatus	purpose
2	Western Kan	2002	summer	kansas	silver iodide, airborne	hail suppress	western kansas
3	Kings River, C	2006	winter	california	silver iodide, ground	airbc precipitation	kings river
4	PGCD Precip	2005	fall	texas	silver iodide, airborne	rainfall enhanc	east central t
5	Panhandle R	2000	summer	texas	silver iodide, airborne	precipitation	panhar
6	2023UTCS-1	2024	winter	utah, nevada	silver iodide, ground	airbc augment	southern and
7	Telluride Ski	2010	winter	colorado	ground	snowpack au upper san	mi
8	Alta and Sno	2012	fa		ground	increase sno alta and	snow
9	Eastern Sierr	2000	winter		airborne	precipitation	eastern sierr
10	Eastern Sierr	2002	winter	california	silver iodide, airborne	precipitation	eastern sierr
11	Eastern San J	2008	winter	colorado	silver iodide, ground	snowpack au eastern san j	
12	2018UTSOC1	2018	winter	utah	silver iodide, ground	augment pre southern	
13	NOAA # 0311	2006	summer	texas	silver iodide, airborne	rainfall enhanc	west texas
14	Telluride Ski	2011	winter	colorado	silver iodide, ground	snowpack au upper san	mi
15	San Joaquin I	2005	winter	california	silver iodide, ground, airbc	precipitation	san joaquin r

Fig. 1 NOAA Form 17-4 Dataset Extraction Pipeline.

Beyond its methodological contributions, the resulting dataset has potential value across a wide range of applications. Researchers can use it to study long-term patterns in weather modification practices, analyze the use and evolution of different seeding agents and deployment methods, or assess geographic and seasonal trends in operational activity.

For example, while there have been several recent research campaigns focused on process-level understanding and evaluation of cloud seeding, such as the 2012 ASCII (AgI Seeding Cloud Impact Investigation) over the Sierra Madre range in Wyoming, or the 2017 Seeded and Natural Orographic Wintertime clouds: the Idaho Experiment (SNOWIE) campaign in Western Idaho^{5,6}, there has been little quantitative analysis of reported cloud seeding activities. With recent renewed interest in both traditional weather modification technology (e.g. the UAE's Rain Enhancement Program and several recent commercial start-ups), as well as geoengineering strategies focused on Solar Radiation Management such as Marine Cloud Brightening that use similar intervention strategies to mitigate climate change⁷, this dataset can provide historical context for weather modification usage in the United States. In addition, this dataset can provide insights into current limitations in reporting requirements that impede large-scale evaluation of the environmental and meteorological implications of weather modification⁴.

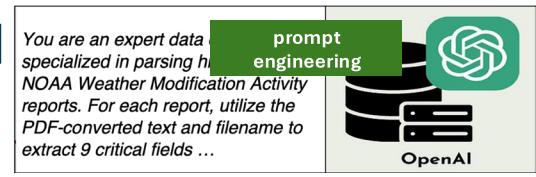
Methods

Input Data Source. Cloud seeding reports were manually downloaded from the NOAA Central Library's Weather Modification Project Reports Archive (<https://library.noaa.gov/weather-climate/weather-modification-project-reports>)³. A total of 1,025 PDF files were collected, including Form 17-4, Form 17-4A, and other supplemental documents. These 1,025 files represent all publicly reported weather modification activities submitted to NOAA. Since the reports are regulatory submissions intended for public access, they are in the public domain and their reuse and redistribution are permitted under U.S. government policy.

Each PDF included Form 17-4, Form 17-4A, or both, along with supplemental pages. Form 17-4 contains the main project summary, including the project purpose, operational period, operator, seeding agents, apparatus for deployment, and target/control areas. Form 17-4A provides tabular data on seeding activity, including days seeded per month and agent quantities released.

For our dataset, we focused exclusively on extracting fields from Form 17-4 which contained the majority of relevant project descriptive information. While many PDFs included supporting materials, such as maps, meteorological analyses, operational logs, and narrative summaries, these were excluded due to inconsistent formatting that would require extensive preprocessing with minimal dataset enhancement.

Preprocessing. A custom Python pipeline was developed to preprocess all Form 17-4 PDF documents (Fig. 1). First, all downloaded files were stored in a local directory. In many cases, a single cloud seeding project was documented across multiple submissions, indicated by suffixes such as .M (initial report), .I (interim report), and .F (final report). These reports were consolidated into a single file per project using a combination of programmatic and manual methods to preserve project-level information. After consolidation, 832 unique weather modification projects remained for processing and inclusion in the dataset.



The scanned documents included both typed and hand-written text, and sometimes a mix of both in the same document. To convert the PDFs into plain text for further processing, a combination of three technologies was used. For native, text-based documents (57% of documents), text was extracted using the `pymupdf` Python library⁸. For scanned or hand-written documents (43% of documents), optical character recognition (OCR) was applied using a combination of `pytesseract`, an open-source wrapper for the Tesseract OCR engine⁹ (processed 16% of all documents), and `l1m-whisperer` (processed 27% of all documents), a layout-preserving PDF extraction technology that combines OCR and native-text extraction¹⁰. All code for the preprocessing pipeline, including file merging, text extraction, and LLM-integration, is publicly available on GitHub (<https://github.com/jdonohue44/NOAA-Weather-Modification-Forms-LLM-Extractor>)¹¹.

In addition to the extracted text, filenames were also provided to the model because the filenames often contained useful contextual clues about the report's location or time frame (e.g. *North Dakota Cloud Modification Project - District I_02-1147_06.01.2002-08.31.2002.pdf*). Notably, there appeared to be two distinct file naming conventions in the archive. Older files (generally before 2017) tend to use a descriptive format that includes the location and date (e.g., *Colorado-SanJuan-2003.pdf*), whereas newer files (from around 2018 onward) follow a standardized alphanumeric format such as *2018TXSWMA-1_2018.pdf*, which embeds the year, state abbreviation, and project code. These filenames were leveraged during extraction to assist with metadata recovery.

LLM Integration. We integrated with OpenAI's platform of LLMs to analyze the extracted text and synthesize the key metadata fields for the structured dataset. After evaluating multiple models, the *o3* model demonstrated the best performance and was selected for all extractions¹².

LLM Prompt. We experimented with multiple prompt templates using prompt engineering techniques to improve extraction quality. The final prompt, shown below with truncation, employs chain-of-thought reasoning by guiding the model through intermediate steps before producing structured outputs. This reasoning style improves performance on tasks requiring multi-step inference and disambiguation¹³. The prompt was used with the *o3* reasoning model for all extractions in the final dataset. All prompt versions and evaluation scripts are available in our public GitHub repository¹¹.

```
# NOAA Weather Modification Report Extraction Expert
You are an expert data extractor specialized in parsing historical NOAA Weather Modification
Activity reports. For each report, utilize the PDF-converted text and filename to extract
**12** critical fields.

## Instructions
For each field, carefully analyze all available information using a step-by-step reasoning
process. Reason step-by-step internally, using evidence from the filename and report content
to resolve conflicting or ambiguous information. Output only the final extracted fields
using the format and rules provided below.

## Fields to Extract
[truncated for brevity]

## Example Reasoning (Internal)
[truncated for brevity]

## Final Extracted Fields Format
Present your final extracted results concisely as follows, in lowercase, comma-separating
multiple values when applicable. Do not include commentary, explanations, or placeholders.
Leave field blank if truly unknowable after exhausting all inference methods using the
filename and text evidence.

PROJECT: [extracted value]
YEAR: [extracted value]
SEASON: [extracted value]
STATE: [extracted value]
OPERATOR AFFILIATION: [extracted value]
AGENT: [extracted value]
APPARATUS: [extracted value]
PURPOSE: [extracted value]
TARGET AREA: [extracted value]
CONTROL AREA: [extracted value]
START DATE: [extracted value]
END DATE: [extracted value]
```

Postprocessing. As instructed, the LLM analyzed the extracted PDF text and responded with a list containing the key information we requested. In order to neatly store this information into CSV format, we then mapped key values into more concise column names, stripped any blank space, and converted to lowercase. Using the formatted data object, each project could be easily written to CSV, with one row per object and columns corresponding to the extracted fields. Lastly, to prepare the final dataset, we used `clean-dataset.py` in¹¹ to remove any duplicates, standardize formatting, and fill blank cells with NaN.

Data Records

The dataset is available on Zenodo as a single CSV file (<https://doi.org/10.5281/zenodo.1492581>)¹⁴. It contains structured information from 832 unique cloud seeding project reports across the United States from 2000 to 2025, with one row per project and one column per extracted field (Fig. 2).

The following fields are present in the dataset:

- **filename:** The original name of the file from which the data was extracted.
- **project:** The official name or designation of the cloud seeding project or activity.
- **year:** The calendar year in which the cloud seeding activity occurred (e.g., 2015).
- **season:** The meteorological or operational season during which the activity took place (e.g., winter, summer).
- **state:** The U.S. state where the cloud seeding operation was conducted (e.g., California).
- **operator_affiliation:** The organization or agency responsible for conducting or overseeing the seeding operation.
- **agent:** The material used to seed clouds (e.g., silver iodide, dry ice).
- **apparatus:** The method or equipment used to deliver the seeding agent (e.g., ground-based generators, aircraft).
- **purpose:** The stated objective of the cloud seeding activity (e.g., augment snowpack, increase rainfall, suppress hail).
- **target_area:** The geographical area where the cloud seeding was intended to take effect.
- **control_area:** A designated area used as a comparison region, typically not subject to cloud seeding.
- **start_date:** The first reported date of seeding activity for the project.
- **end_date:** The final reported date of seeding activity for the project.

Data Visualizations. Between 2000 and 2025, cloud seeding activity in the United States was geographically concentrated in western states that rely on snowpack for water supply. California, Colorado, and Utah accounted for the majority of seeding activities, with summertime rain enhancement in Texas also contributing a substantial number of activities (Fig. 3). These records, shown in Fig. 4, align with states that maintain active cloud seeding programs⁴.

The primary stated purpose of cloud seeding events was to increase snowpack, followed by increasing precipitation and rain (Fig. 5). Snowpack enhancement is particularly important in mountainous regions where runoff from melting snow supplies reservoirs and irrigation systems.

Silver iodide is by far the most common seeding agent, especially in ground-based and airborne operations (Fig. 6). Other agents such as calcium chloride and carbon dioxide are used less frequently. Ground deployment is the most prevalent deployment method, particularly when using silver iodide, although a mix of ground and airborne (ground, airborne) approaches is also present.

Over this 25-year period, the number of weather modification events peaked in the early to mid-2000s, then declined through the 2010s, before rebounding in 2024 and 2025 (Fig. 4).

Technical Validation

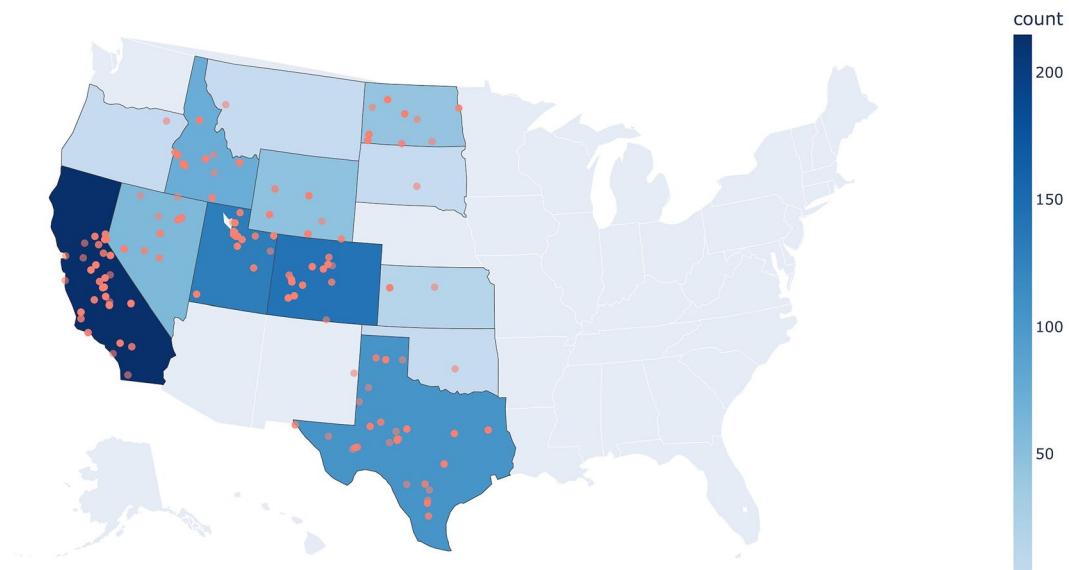
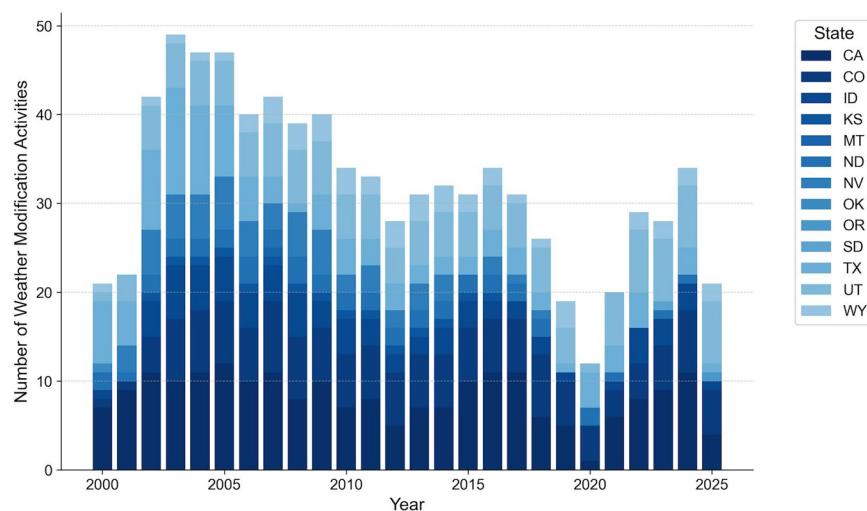
To address the potential for LLMs to hallucinate content not present in the source documents, and the inconsistent quality of the input PDFs themselves, we measured field-level extraction accuracy by manually reviewing a random sample of 200 records extracted by the LLM and comparing them to ground truth. This evaluation provides an empirical bound on potential extraction errors in the released dataset, enabling downstream users to incorporate measurement uncertainty into their analyses.

Although LLMs are capable of producing fluent and structured text, they can generate information that is not present in the source material¹⁵. In the context of this paper, such hallucinated content could include incorrect dates, locations, or project descriptions, which would compromise the reliability of downstream analyses involving historical weather modification activities in the United States.

Additionally, the quality of the input documents presented another challenge. Many PDFs contained handwritten annotations, low-resolution scans, or incomplete text referring the reader to information from a previous year or external attachments rather than providing complete answers in the designated fields. For example, reports from the North Dakota Cloud Modification Project and the Western Kansas Weather Modification Program left fields blank and instead referred to separate documents such as an operational plan or an attached sheet. These references could not be processed directly, limiting the ability of the model to extract a full set of structured values.

Extraction Accuracy. From the population of $N=832$ records we randomly sampled $n=200$ to evaluate field-level extraction accuracy (Table 1). This sample size was chosen to yield a margin of error of ± 3 percentage points at the 95% confidence level. The evaluation was conducted by two independent human experts manually creating a golden dataset (ground truth) using the source PDFs, then comparing the extracted fields against the golden dataset using `compare-to-golden.py`¹¹.

A	B	C	D	E	F	G	H	I	J	K	L	M	
1	filename	project	year	season	state	operator_affiliation	agent	apparatus	purpose	target_area	control_area	start_date	end_date
2	2016CAKASF	kaweah cloud	2017	winter	california	rhs consulting ltd	silver iodide, airborne	augment snowpack		kaweah river basin	merced river, 11/1/16	4/30/17	
3	2016CAKESP	kern river clo	2017	winter	california	rhs consulting ltd	silver iodide, airborne	augment snowpack		kern river basin	merced river, 11/1/16	4/30/17	
4	2017CAKASF	kaweah cloud	2018	winter	california	rhs consulting ltd	silver iodide, airborne	augment snowpack		kaweah river basin	merced river, 11/1/17	4/30/18	
5	2017CAKESP	kern river clo	2018	winter	california	rhs consulting ltd	silver iodide, airborne	augment snowpack		kern river basin	merced river, 11/1/17	4/30/18	
6	2017IDCCSN	clark county	2018	winter	idaho	clark county	silver iodide	ground, airt	augment snowpack	clark county, idaho		11/1/17	5/15/18
7	2018CAKRAV	kings river clo	2019	winter	california	north american weat	silver iodide	ground, airt	increase precipitation	kings river watershed	various sites	11/15/18	3/31/19
8	2018CASBSL	santa barbara	2019	winter	california	north american weat	silver iodide	ground	increase precipitation	santa barbara and various sites		11/15/18	4/15/19
9	2018CAUSJS	san joaquin ri	2019	winter	california	desert research insti	silver iodide	ground	augment snowpack	south fork san joaquin river basi		11/20/18	6/20/19
10	2018CAUTOL	upper tuolumne	2019	winter	california	weather modification	silver iodide	airborne	increase precipitation	upper tuolumne river watershed		11/6/18	4/30/19
11	2018COCMR	central color	2019	winter	colorado	western weather cor	silver iodide, ground		augment snowpack	eastern central co	northern colorado	11/1/18	4/15/19
12	2018COESJM	eastern san ju	2018	winter	colorado	western weather cor	silver iodide, ground		augment snowpack	eastern san juan n	northern new	11/1/18	4/15/19
13	2018COPWS	purgatory/we	2019	winter	colorado	western weather cor	silver iodide, ground		augment snowpack	western san juan r	se utah mour	11/1/18	3/31/19
14	2018COTUSI	telluride ski a	2019	winter	colorado	western weather cor	silver iodide, ground		augment snowpack	upper san miguel i	uncompahgr	11/15/18	4/15/19
15	2018COBV	central color	2019	winter	colorado	western weather cor	silver iodide, ground		augment snowpack	vail and beaver cre	flat top mour	11/1/18	3/31/19

Fig. 2 Screenshot of the CSV Dataset (*cloud_seeding_us_2000_2025.csv*).**Fig. 3** Cloud Seeding Activity by U.S. State (2000–2025). States with active weather modification programs show the highest number of recorded operations. Specific locations (shown as salmon markers) are from geocoding the stated *target_area* field with the GoogleMaps API.**Fig. 4** Cloud Seeding Activity by U.S. State over Time (2000–2025). The number of activities peaked between 2003–2005, declined gradually, and rose again after 2021.

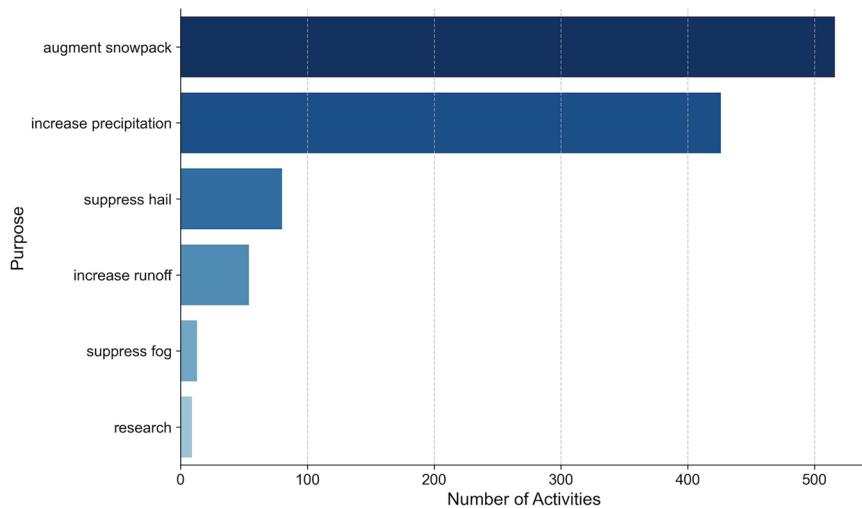


Fig. 5 Stated Purpose of Cloud Seeding Activity (2000-2025). Augmenting snowpack is the leading purpose, followed by increasing precipitation.

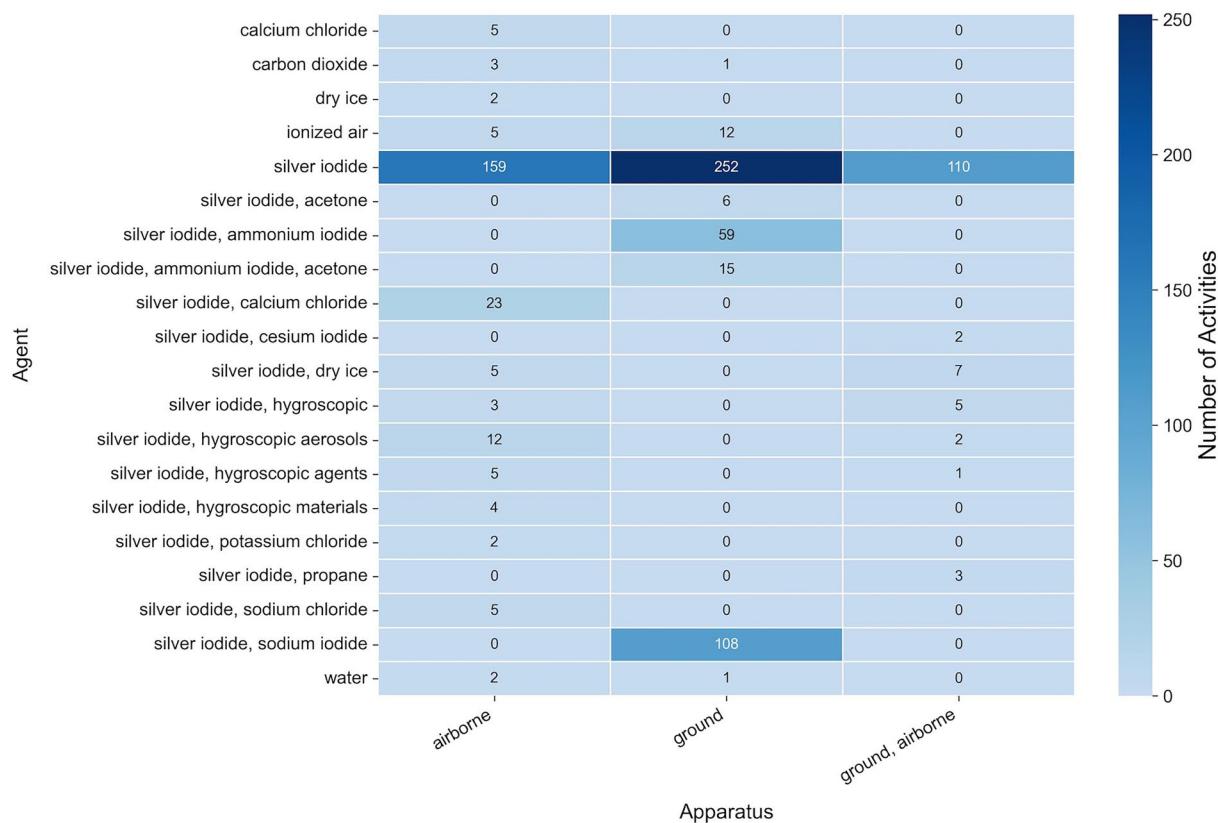


Fig. 6 Agent and Apparatus Used for Cloud Seeding Operations (2000-2025). Silver iodide dominates among agents and is deployed most frequently using a ground-based apparatus.

LLM Comparisons. We evaluated several language models to improve extraction accuracy, including *gpt-4.1*, *gpt-4.1-mini*, *gpt-4.0-mini*, *o4-mini*, and *o3*¹² (Table 2). To isolate the effect of model choice, we held the LLM prompt constant using a fixed prompt version (Prompt C¹¹) across all evaluations. For each trial, model outputs were compared to human-labeled values across all fields, and the proportion of correct fields was averaged across the dataset.

The *o3* reasoning model achieved the highest overall accuracy at 96.33%, but was initially an order of magnitude more expensive than *o4-mini* costing approximately \$0.05 per document versus \$0.005 for *o4-mini*, which achieved 95.00% accuracy. However, in June 2025, OpenAI reduced the price of *o3*, bringing its cost down to \$0.01 per document. Following this price drop, we adopted *o3* to maximize accuracy while maintaining low cost.

Field	Accuracy (%)	95% Confidence Interval
Project	100.00%	[97.00%, 100.00%]
Year	100.00%	[97.00%, 100.00%]
Season	100.00%	[97.00%, 100.00%]
State	99.50%	[96.50%, 100.00%]
Operator Affiliation	98.50%	[95.50%, 100.00%]
Agent	98.50%	[95.50%, 100.00%]
Apparatus	97.00%	[94.00%, 100.00%]
Purpose	99.50%	[96.50%, 100.00%]
Target Area	97.50%	[94.50%, 100.00%]
Control Area	92.00%	[89.00%, 95.00%]
Start Date	98.50%	[95.50%, 100.00%]
End Date	99.50%	[96.50%, 100.00%]
Overall Average	98.38%	[95.38%, 100.00%]

Table 1. Field-Level Extraction Accuracy (n = 200; LLM = *o3*).

LLM	Accuracy (%)	95% Confidence Interval
gpt-4.1	93.33%	[86.11%, 97.48%]
gpt-4.1-mini	93.33%	[86.11%, 97.48%]
gpt-4o-mini	91.67%	[84.03%, 96.34%]
<i>o3</i>	96.33%	[88.08%, 98.67%]
<i>o4-mini</i>	95.00%	[88.08%, 98.67%]

Table 2. Extraction Accuracy by LLM (n = 50; prompt C).

Prompt	Accuracy (%)	95% Confidence Interval
Prompt A	93.67%	[86.64%, 97.56%]
Prompt B	44.00%	[30.98%, 57.69%]
Prompt C	95.00%	[88.08%, 98.67%]

Table 3. Accuracy by LLM Prompt (n = 50; LLM = *o4-mini*).

Prompt Engineering. To isolate the effect of prompt design, we measured extraction accuracy with a single model, *o4-mini* (Table 3). We chose *o4-mini* because it offers a favorable balance between cost and performance. While the absolute accuracies are specific to *o4-mini*, we expect the relative ranking of prompts to extend to other models, such as *o3*.

Three prompt formats were tested: detailed markdown instructions (Prompt A), concise instructions (Prompt B), and a chain-of-thought format (Prompt C) that encouraged step-by-step reasoning. Prompt C achieved the highest accuracy at 95.00%, suggesting that explicit reasoning steps can substantially improve structured data extraction.

Prompt B performed poorly (44.00%) due to frequent missing values in the outputs, likely caused by its lack of field-level guidance and examples. In contrast, Prompt A performed better (93.67%) and benefited from including worked examples for each field. While not explicitly labeled as chain-of-thought, Prompt A was influenced by the same principles, demonstrating that even partial integration of reasoning structure can enhance accuracy.

Together, these validation efforts show that LLM-based data extraction accuracy can be significantly improved through deliberate prompt design and model selection. In particular, prompts incorporating chain-of-thought reasoning and field-specific examples consistently enhanced extraction accuracy, even in complex historical documents with heterogeneous formatting and quality. These findings suggest that careful prompt engineering should be considered a core methodological step when using LLMs to extract scientific data from historical documents.

Usage Notes

All weather modification reports in this dataset are self-reported by project operators and may not represent a complete record of U.S. weather modification activity, particularly in cases of unsubmitted or omitted filings. In addition, research field campaigns that piggy-backed on existing cloud seeding operations, such as the ASCII field campaign's monitoring of the Wyoming Weather Modification Pilot Project⁵, may not show up as distinct entries in the data record.

As of this writing, NOAA's publicly available records span the years 2000-2025. If older reports (from 1972-1999) are recovered and released, the current extraction pipeline can be applied with minimal modification to extend the dataset's temporal coverage.

For PDF-to-text conversion, note that `llm-whisperer` is free only for up to 100 PDF pages per day. Additionally, note that the OpenAI models referenced (e.g., *gpt-4.1*, *gpt-4.1-mini*, *gpt-4o-mini*, *o4-mini*, and *o3*), incur usage costs depending on model and volume, so users should consult current OpenAI pricing for budgeting larger-scale replication¹⁶.

Data availability

The dataset is publicly available on Zenodo (<https://doi.org/10.5281/zenodo.14925811>)¹⁴. The original NOAA files used as input are publicly available on NOAA (<https://library.noaa.gov/weather-climate/weather-modification-project-reports>)³.

Code availability

Python code for extracting the dataset from the NOAA reports can be obtained from our public GitHub repository (<https://github.com/jdonohue44/NOAA-Weather-Modification-Forms-LLM-Extractor>)¹¹. Python code for visualizing the dataset is also publicly available on GitHub (<https://github.com/jdonohue44/US-Cloud-Seeding-Analysis>)¹⁷.

Received: 14 May 2025; Accepted: 3 November 2025;

Published online: 11 December 2025

References

1. Desert Research Institute, "Making it snow: A brief history and review of the science behind cloud-seeding," Accessed: (2025-07-21) (2023).
2. National Oceanic and Atmospheric Administration, "Code of Federal Regulations, Title 15, Part 908: Maintenance of Records and Submission of Reports on Weather Modification Activities." <https://www.ecfr.gov/current/title-15/subtitle-B/chapter-IX/subchapter-A/part-908>, Accessed: 2025-04-01 (2024).
3. NOAA Central Library "Weather Modification Project Reports Archive." <https://library.noaa.gov/weather-climate/weather-modification-project-reports>, Accessed: 2025-04-01 (2025).
4. U. S. G. A. Office, "Cloud Seeding Technology: Assessing Effectiveness and Other Challenges | U.S. GAO."
5. Geerts, B. "The Ag1 seeding cloud impact investigation (ASCI) campaign 2012: Overview and preliminary results," *The Journal of Weather Modification* **45**, 24–43 (2013).
6. Tessendorf, S. A. "A Transformational Approach to Winter Orographic Weather Modification Research: The SNOWIE Project," *Bulletin of the American Meteorological Society* **100**, 71–92 (2019).
7. Feingold, G. *et al.* "Physical science research needed to evaluate the viability and risks of marine cloud brightening." *Science Advances* **10**, eadi8594 (2024).
8. Hansen, N., "pymupdf4llm: Structured PDF text extraction for LLMs," Accessed: 2025-04-01 (2024).
9. Smith, R. "An Overview of the Tesseract OCR Engine," *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)* **2**, 629–633 (2007).
10. Zipstack, I., "LLM Whisperer Python Client." https://docs.unstract.com/llmwhisperer/1.0.0/llm_whisperer/python_client/, Accessed: 2025-04-20 (2025).
11. Donohue, J., "NOAA Weather Modification Forms LLM Extractor." <https://github.com/jdonohue44/NOAA-Weather-Modification-Forms-LLM-Extractor>, Accessed: 2025-05-01 (2025).
12. OpenAI, "OpenAI models documentation." <https://platform.openai.com/docs/models>, Accessed: 2025-04-01 (2025).
13. Wei, J. *et al.* "Chain-of-thought prompting elicits reasoning in large language models." *Advances in Neural Information Processing Systems 36 (NeurIPS 2022)*. Curran Associates, Inc. <https://doi.org/10.5555/3600270.3602070> (2022).
14. Donohue, J. J. and Lamb, K. D., "Cloud seeding activities in the United States (2000–2025)." <https://doi.org/10.5281/zenodo.14925811> (2025).
15. Huang, L. *et al.* "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems* **43**, 1–55 (2025).
16. OpenAI, "API pricing." <https://openai.com/api/pricing/> Accessed: 2025-04-22 (2025).
17. Donohue, J. "US Cloud Seeding Analysis." <https://github.com/jdonohue44/US-Cloud-Seeding-Analysis>, 2025. Accessed: 2025-05-01.

Acknowledgements

Both authors acknowledge funding from Columbia University's Data Science Institute. K.D.L. acknowledges funding from the Zegar Family Foundation.

Author contributions

J.J.D. designed the extraction pipeline, LLM integration, evaluation framework, and led manuscript writing. K.D.L. designed the research and acquired funding for the research. K.D.L. also supervised the project and co-wrote the manuscript. All authors contributed to the data visualizations. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.J.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025