

Prospects for Machine Learning Activity within the United States National Weather Service

Paul J. Roebber  and Stephan Smith

KEYWORDS:

Atmosphere;
North America;
Forecasting;
Machine learning

ABSTRACT: The National Weather Service (NWS) Office of Science and Technology Integration commissioned a report to assess the status of artificial intelligence (AI) and machine learning (ML) activity within the agency with a view toward identifying existing obstacles and recommending future directions. The purpose of this essay is to communicate the steps that the NWS plans to take to realize the potential benefits of AI in operations. AI activities are growing rapidly within atmospheric sciences, and the NWS is part of this growth. However, the activity is fragmented and lacks the needed infrastructure for improved coordination of effort. Current obstacles to progress include insufficient workforce training in AI/ML, a lack of curated datasets and software that can be used for development and evaluation of these approaches, the absence of a centralized clearing house available to NWS personnel for technical expertise and consultation, limited operational compute resources, and a lack of a clear end-to-end project pathway that encompasses exploration, development, testbed/proving ground, and operational implementation. These limitations are addressable. Training materials specific to NWS interests can be developed through collaboration with existing NOAA centers. Establishing a reference library staffed with AI/ML consultants tasked with collaborating with operational units would reduce siloed efforts and enhance productivity. Establishing funding vehicles for theme-based projects with a sustainable pathway through operational implementation would help bridge the research-to-operations "valley of death." Given the growth of AI/ML across the U.S. Weather Enterprise and the already substantial involvement of academic and private sector entities, these developments within the NWS will be of interest to the atmospheric science field.

<https://doi.org/10.1175/BAMS-D-22-0181.1>

Corresponding author: Paul J. Roebber, roebber@uwm.edu

In final form 17 May 2023

©2023 American Meteorological Society

For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](#).

AFFILIATIONS: Roebber—Atmospheric Science Program, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin; Smith—Office of Science and Technology Integration, NOAA/National Weather Service, Silver Spring, Maryland

This essay summarizes the findings of a report commissioned by the National Weather Service (NWS; see Table 1 for acronym definitions) Office of Science and Technology Integration to evaluate prospects for the agency in the artificial intelligence (AI)/machine learning (ML) domain [see Roebber (2022) for full details]. The intent of this essay is to communicate the steps that the NWS plans to take to realize the potential benefits of AI in operations. The relevance to the wider atmospheric science community is the involvement of weather service partners, including those in academia, the private sector, and NOAA laboratories. This effort is aligned with the Priorities for Weather Research report (hereafter PWR-2021; NOAA Science Advisory Board 2021), which states a need to “target the understanding and prediction of high-impact weather to match the urgent need imposed by climate trends, population and infrastructure increases, and disproportionate impacts on vulnerable communities; including exploring new innovations with AI and machine learning applications.” Similarly, the 2020 NOAA Artificial Intelligence Strategy (NAIS) states a vision that the “expansion of Artificial Intelligence (be) accelerated across the entire agency to make transformative improvements in NOAA mission performance and cost effectiveness.”

In performing this evaluation, it is necessary to consider some context. Several definitions and understandings of ML and its distinction from AI exist. For example, from the NAIS: “Artificial Intelligence refers to computational systems able to perform tasks that normally require human intelligence, but with increased efficiency, precision, and objectivity. A subset of AI called machine learning refers to mathematical models able to perform a specific task without using explicit instructions, instead relying on patterns and inference.”

Table 1. A list of acronyms.

Acronym	Definition	Acronym	Definition
AI	Artificial intelligence	NAIS	NOAA artificial intelligence strategy
ANN	Artificial neural network	NBM	National blend of models
CNN	Convolutional neural network	NCAI	NOAA’s Center for Artificial Intelligence
COVID	Coronavirus disease	netCDF	Network Common Data Form
CPU	Central processing unit	NOAA	National Oceanic and Atmospheric Administration
CSTAR	Collaborative Science, Technology, and Applied Research	NSF	National Science Foundation
GPU	Graphics processing unit	NWP	Numerical weather prediction
HPC	High performance computing	NWS	National Weather Service
JTTI	Joint Technology Transfer Initiative	PWR-2021	NOAA Science Advisory Board (2021)
MDL	Meteorological Development Laboratory	RF	Random forest
ML	Machine learning	RL	Readiness level
MLP-ANN	Multilayer perceptron ANN	R2O	Research-to-operations
MLR	Multiple linear regression		

To some degree, this definition overstates what is being accomplished with present technologies. To be sure, evolutionary history has ingrained the importance of pattern recognition as an element of human intelligence, including the tendency toward overforecast bias (e.g., Foster and Kokko 2008; Shermer 2008)—but true intelligence cannot be divorced from context. Ecologically speaking, the mind is gauging value, which is context dependent (Barrett 2021). For example, whether it is worth expending energy to obtain food depends on both the current and future state (hunger) and how much energy expenditure is needed. Our brains are prediction engines, trying to anticipate what sensory inputs they will receive next. Presently, AI/ML is not capable of this form of generalization, although generalized AI is an active area of research (McKenna et al. 2020).

Instead, AI/ML approaches today are based upon “learning” patterns in a narrow and rigorously defined framework, such as skill games like chess or Go. We have used quotation marks above, since whether or not an ML algorithm truly learns is a matter of debate and subject to the specific definition of learning (e.g., Kodelja 2019). Some of the most powerful chess engines now use the form of ML known as reinforcement learning, which essentially instructs the computer to play hundreds of millions of games against itself, building its expertise through these trials. In so doing, the algorithm finds the most optimal moves in the most probable situations. The most successful chess grandmasters look for lines of play that are useful but have not necessarily been favored in computer evaluations. As noted by Litt (2020), human and chess engine *teams* are often superior to humans or machines alone, and these teams do not necessarily include top-ranked human players. This is reminiscent of the experienced forecaster who uses computer guidance to inform but not supplant personal judgment and reflects the critical importance of understanding the strengths and limitations of the automated guidance. Trustworthiness and presumably the quality of the resulting forecast will depend on that understanding (see below for more discussion of AI/ML trustworthiness).

There is a long history of weather forecasters “teaming” with guidance to perform better than human forecasters or guidance alone, as evidenced by the added skill of NWS forecasters relative to guidance [e.g., Fig. 5 of Schaffer et al. (2020) for tropical storms]. However, the forecast itself is only one part of the forecast process. The communication of forecaster understanding of the weather risks specific to those partners exposed to that risk or tasked with working with those exposed to the risk (such as emergency managers) is the “last mile” of the forecast process (e.g., Lazo et al. 2016). AI/ML reflects a potential acceleration of the shift in NWS focus from forecast generation to weather decision support, since better guidance tools can free time for forecasters to focus more on those interactions with core partners.

In the atmospheric sciences, data analysis and postprocessing have been rooted in traditional statistical methods. The most noteworthy and longstanding example of this is model output statistics (MOS; Glahn and Lowry 1972), which uses the well-established method of multiple linear regression (MLR) to produce forecast variables in an operational context. In operations, the focus is necessarily on making predictions. While this focus seems particularly well suited to a field organized around forecasting, the physics basis of atmospheric science argues for a need to understand as well as predict, and the confidence that ensues from understanding *why* a prediction is being made. One advantage of linear methods like MLR is the relative ease in understanding the relationship between the inputs and the prediction—there is a trade-off between the additional skill obtainable in some forecast problems by accounting for nonlinearities as directly achievable through AI/ML (e.g., neural networks) and the increased difficulty in understanding the result obtained using those methods.

Further, it is now recognized that the so-called objective nature of AI/ML does not imply a lack of bias, since these techniques fundamentally depend on the choices of the developers in terms of what data to use, what metrics define success, and the fundamental limitations of the data itself (quantity, quality, collection patterns, etc.). This issue will become increasingly

important as social science applications become more common in the decision support context of weather forecasting.

All the above suggests that AI/ML algorithms should be considered another useful tool for informing the decision-making process, but in making those decisions, a deep understanding of both the strengths and weaknesses (such as inherent biases) of these algorithms is essential. This speaks to the need for a trained workforce—not necessarily algorithm developers but coordinated efforts between those developers and domain experts who are tasked with using these tools in the decision support process. These perspectives inform the contents of this essay, which was assembled with contributions from a large cohort of NWS practitioners and external (academic) collaborators, whose collective work involves research, implementation, and application of these and other forecast tools. We also draw from Haupt et al. (2021) and Boukabara et al. (2021), the latter of which reports on a 2019 workshop that brought together over 400 scientists, program managers, and leaders from the public, academic, and private sectors involved in the development and adaptation of AI tools and applications.

Current activities

AI/ML has a long history within the environmental sciences (Haupt et al. 2022). Attendees of recent annual meetings of the American Meteorological Society (AMS) have qualitatively experienced a rapid growth in community data science activity, as indicated by the number of papers presented, integrated throughout multiple sessions (Fig. 1) [see also Fig. 1 from Chase et al. (2022) for longer-term trends in AI/ML publications].

NOAA's Center for Artificial Intelligence (NCAI) provides a count of data science projects across the agency, and this count also reveals considerable activity by NWS and across NOAA's Line Offices and mission areas, with ~188 self-reported projects in 2020 and ~263 in 2022 (R. Redmon 2022, personal communication). Indeed, in response to this growing interest in the field, the AMS recently launched a new journal devoted to AI/ML for Earth systems (www.ametsoc.org/index.cfm/ams/publications/journals/artificial-intelligence-for-the-earth-systems/).

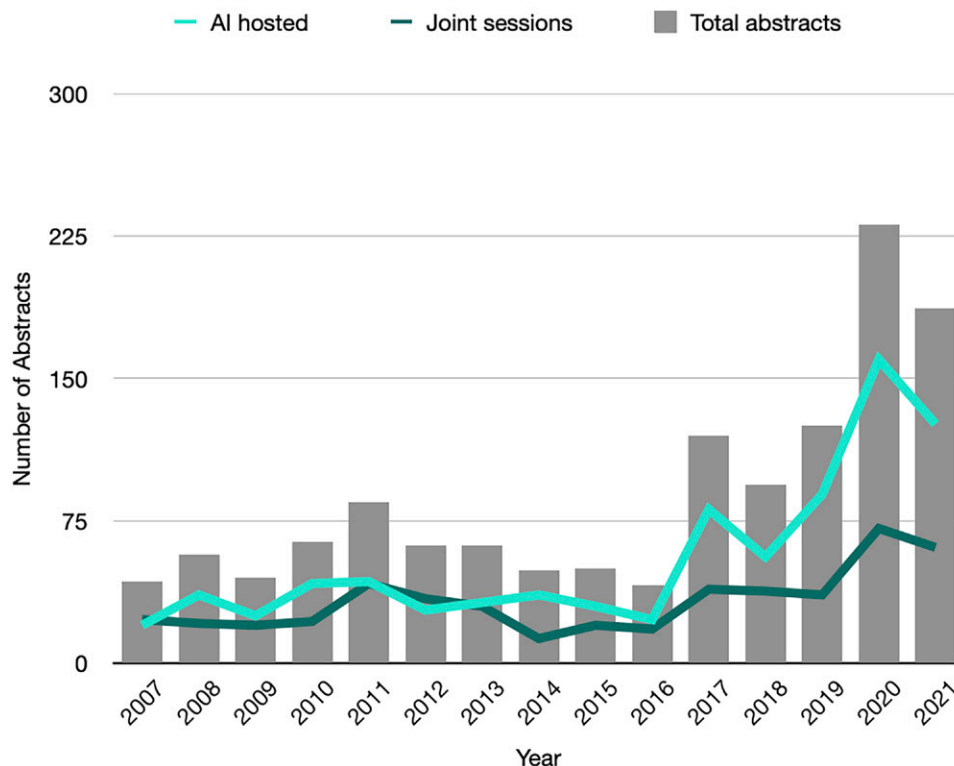


Fig. 1. AI/ML activity (abstracts presented, AI and joint sessions held) at the AMS Annual Meeting (A. McGovern 2022, personal communication).

A number of AI/ML projects are currently underway within the NWS. A non-exhaustive list providing some sense of the variety of these projects and their distribution across the NWS is provided in Roebber (2022). This set of projects mirrors the findings of the NAIS, which listed a number of existing AI/ML efforts within NOAA that pertain to the NWS, including 1) quality control of weather observations; 2) improving physical parameterization for weather, ocean, ice modeling, and improving the computational performance of numerical models; 3) aiding weather warning generation; 4) supporting partners in wildfire detection and movement; and 5) using machine learning for reliable and efficient processing, interpretation, and utilization of Earth observations. We note that AI/ML tools, if implemented properly, also can assist operations with the well-known problem of data overload, since trained users can deploy them to get the sense of large volumes of data and extract explicit information relatively quickly. Given this potential, several “wish list” projects were also identified by contributors to this review (Roebber 2022). A subset of these includes week two extreme events, grid and scenario-based postprocessing, data-driven water reservoir predictions, and computationally efficient data extraction from numerical weather prediction (NWP) models.

At this stage, one can view NWS AI/ML activity as broad-based and growing, but uncoordinated. Again, from the NAIS: “Despite this notable progress, the true potential for AI to advance NOAA’s mission has not been realized because all NOAA AI activity heretofore has originated within individual offices with no institutional support. Additionally, some development has been redundant because of a lack of awareness across the agency due to the absence of a coordinating directive or authority.” Our survey suggests the roots of this redundancy are as noted above, but also owing to the need to perform any such coordination with a thoughtful inclusion of the expertise and operational requirements of specific entities (e.g., the needs of the Storm Prediction Center are not identical to those of the National Hurricane Center). Accordingly, any proffered solutions must take this needed domain expertise and site-specific application into account.

Limitations to future progress

Successful ML development depends on three pillars: 1) ample, quality-controlled datasets; 2) technical skills for development; and 3) domain expertise—familiarity both with the forecast problems and the operational logistics of the setting where that problem is being considered. Several roadblocks in the path from research-to-operations (R2O) exist in the AI/ML domain, relevant to each of these pillars, and are discussed in turn below.

Workforce training and domain expertise. The PWR-2021 report noted an important workforce challenge related to AI/ML, specifically “staying nimble requires a workforce with a broader and evolving range of technical skills and spectrum of talents. Future workforces will include meteorologists working with other experts in Earth sciences, high performance computing (HPC), artificial intelligence (AI) and machine learning (ML), observing, data assimilation, modeling technologies, social sciences, etc. Strategies to increase the workforce capacity will be essential given the increasing demands for these skills.”

Currently, NWS employees tend to come from two main areas of academic training, neither of which explicitly requires training in AI/ML. While the physical scientist classification (GS-1301) allows some flexibility, the GS-1340 requirements for meteorologists leave little room for expansion and are now decades old. It is not necessarily the case that every NWS employee must be an AI/ML expert with the ability to develop such models themselves, but at a minimum, they should be sufficiently aware of this domain to speak intelligibly with AI/ML developers and to recognize both opportunities and pitfalls associated with application of these technologies to areas within their purview. Accordingly, the NOAA/NWS standards

should be revised to promote greater flexibility and expansion of skills that are needed now and those that will be needed in the future (e.g., Stuart et al. 2022).

This reinforces the need for domain expertise. Since the technical expertise for AI/ML is substantial, often such experts come from computer science, and such individuals are typically not well versed in the details of either the meteorology or operational logistics. While it is possible to come from such a background and gain additional knowledge through further academic training (e.g., coursework) and work experience, deep understanding of operational needs and logistical challenges in addition to meteorological knowledge are required to build AI/ML tools that can be incorporated into routine use. This suggests that a balance between centralization and local expertise needs to be established for successful coordination of AI/ML activity across the NWS.

Beyond the development stage, there is a need for users to be sufficiently aware of the strengths and weaknesses of AI/ML tools to use them judiciously rather than simply as a black box. Equally as important, a lack of sophisticated understanding of these strengths and weaknesses likely reinforces resistance to change rather than an attitude of exploring possibilities. This connects to the active areas of interpretable ML and trustworthiness—for such tools to be employed, their credibility will be critical (see also Boukabara et al. 2021).

The National Artificial Intelligence Research and Development Strategic Plan (NSTC 2019) emphasized the development of trustworthy AI systems. The National Science Foundation (NSF) call for proposals for National Artificial Intelligence (AI) Research Institutes (<https://www.nsf.gov/pubs/2022/nsf22502/nsf22502.htm>) notes that technologies are trusted “because they are reliable, predictable, governed by rigorous and measurable standards, and provide the expected benefits.” Cains et al. (2023) explored trustworthiness in the context of weather forecasting by studying the use of AI/ML tools by a set of 16 NWS forecasters from different regions. They found that forecaster understanding of the guidance involves understanding the functionality of the model, its strengths and weaknesses, how the model performs under different scenarios, and how the model performs compared to other guidance. Further, Cains et al. (2023) emphasize that forecasters need personal experience with the guidance, that is, to be able to use the tool both before and during use, since this allows the forecasters to interact and interrogate the proffered solutions, and to develop mental bias corrections of the model’s performance. We expect that, if such experience is developed, the resulting sophisticated application of AI/ML tools will benefit weather decision support efforts.

Data and computational resources. Data requirements for both traditional and AI/ML techniques are extensive. In the case of AI/ML, to develop such algorithms, the best practice is to split these data into three segments: a training segment, a validation segment, and an independent test segment. The training segment is used for model development (e.g., to tune the weights and biases of a neural-network model); the validation segment is used to select model hyperparameters (e.g., the optimal number of layers or weights for a neural network); and the test segment is used to evaluate the generalization of the results once training and tuning are completed.

Datasets are necessarily large, since this process requires “exploration” of the n -dimensional variable space—if these data do not sufficiently fill this space (i.e., are not comprehensive nor representative, as might be the case for an extreme event), then the AI/ML scheme may not be able to produce a good mapping of inputs to outputs in that area, leading to potential performance errors (see also Boukabara et al. 2021). Further, since this mapping is nonlinear, multiple examples within a data neighborhood help to reduce the deleterious impacts of noisy data, further increasing the needed size of the dataset.

Second, it is a truism of this work that considerable time is spent simply managing datasets (aka data wrangling). For example, in a survey conducted by the Earth Science Information Partners Data Readiness cluster (an open cross-sector collaboration that the NCAI contributes to), responses to the question “In your typical AI/ML application development, roughly what percentage of your time do you spend on finding, accessing, and preprocessing data?” showed that only 20% of the survey group spent a quarter or less of their time on that activity, and nearly half indicated that they spent the majority of their time on this task (R. Redmon 2022, personal communication). This is the case since the variety of needed inputs come in many different formats, from multiple sources, and may further need to be synced in space and time before being presented to an AI/ML scheme for training. These data must also be quality controlled to limit the amount of noise that is presented. The consequence of these requirements is that some problems are not amenable to AI/ML algorithm development, either because the data needed do not exist or the time requirements for the development task are too extensive.

Space and compute needed for development of AI/ML tools likewise can be substantial, owing to the size of the datasets and the data cycling needed for training those algorithms. Further, such training can be more effective when graphics processing units (GPUs) rather than centralized processing units (CPUs) are available, owing to their ability to process large blocks of data in parallel. Currently, such a development system is lacking. Operational computational resources are also a limitation, since finding compute slots on the NWS operational system is always a challenge. Without increased availability of these computational resources, development of AI/ML tools will be constrained and when developed, the transition of those technologies to operations will not occur (see also PWR-2021).

Fundamental AI/ML research. There is a need for exploratory AI/ML work, which contrasts with the readiness level (RL) criteria used in collaborative opportunities such as the Joint Technology Transfer Initiative (JTTI) program. The JTTI program requires RL 4 or above, defined as a concept that has been already developed and validated and is ready to be tested in a NOAA pseudo-operational environment. In the AI/ML domain, owing to the rapid development in this field and the need to do considerable exploration of new approaches in an operational context, the limited opportunities for funding research at lower readiness levels blocks innovation. This approach promotes incremental rather than the high-risk–high-reward work that is needed.

One academic contributor commented that there are insufficient dedicated resources to increase RLs on projects, for example, the difficulty of showing that the technique works well enough to gain review in an operational testbed. While NOAA’s Oceanic and Atmospheric Research laboratories are intended to be the places where mid-RL level research is done within NOAA, there is a lack of coordination between efforts across the agency. Another substantial obstacle is the inconsistent timeline between academic work and operations.

Whether or not such tools are developed and have potential to be operationally useful, transitioning them to the operational computing system (such as NOAA’s Weather and Climate Operational Supercomputing System or a cloud system) is a further challenge, owing to availability of that resource (such as limited compute slots or funding for time on a cloud system) and the time of NOAA collaborators to effect implementation and support. At present, there is, in this sense, no centralized home or dedicated support for AI/ML development within the NWS.

More generally, there are relatively few funding opportunities for academic collaborators, and as such, even those academics inclined to pursue the difficult and time-consuming work of bridging the R2O gap are often better able to succeed professionally by directing their efforts in more traditional ways, such as fundamental research through NSF grants. This R2O “valley of death” is not specific to AI/ML alone but is simply one example of more widespread

organizational issues within NOAA (NWP is another example). Regardless, from an AI/ML standpoint, it is important to note that this is a consequence of the misalignment between program structures (the number of programs for which funding is available, the amount of funding, and the time periods over which the awarded work is to be conducted) and academic requirements for research. This means that the valley can be crossed if efforts are undertaken to do so which would be broadly beneficial to the U.S. Weather Enterprise.

Recommendations

There are potential solutions to each of the above obstacles. Naturally, each of these solutions is subject to particulars regarding future funding and staffing, for which we have no foreknowledge and do not make any predictions.

Workforce training and domain expertise. It would seem reasonable to identify the NCAI as an AI/ML training resource. This group has already begun efforts to develop example Jupyter notebooks and R materials organized in a “learning journey” style to ultimately encourage the broad AI community of practice to contribute materials to an NCAI curated library. Additional materials specific to NWS interests could be developed with collaboration from National Centers and other NWS entities [e.g., postprocessing with the Meteorological Development Laboratory (MDL)], and NCAI staff have indicated an interest in undertaking that effort. In that regard, NCAI has requested approval for a public repository landing page (NOAA GitHub), where they would stage NCAI created and contributed examples (e.g., rip current detection and others are in development) and expand from there with contributions across NOAA. It is likely that a focus on hiring in the NWS with scientific background in both meteorology and AI/ML will be needed in addition to enhancing the training of existing staff. Beyond providing training materials, workforce training and development can be accelerated by improved coordination of AI/ML activities between product developers and forecasters. This topic is explored below.

Data and computational resources. The current uncoordinated nature of workforce training could be ameliorated with a software, data, and consulting clearing house or library (e.g., Fig. 2 of Hamill 2015). This library would include a variety of standardized datasets that could be used to develop different types of AI/ML applications, depending on the need, and most importantly, as a reference against which to compare AI/ML applications. Notably, this echoes the recommendations of Haupt et al. (2021), who initiated the development of an open-access experimental testbed database containing five datasets, as well as code to aid in rapid analysis and evaluation of results. Dueben et al. (2022) extend that work by providing a definition of benchmark datasets for weather and climate applications and a list of the benchmark datasets that will be needed.

This library could include modular software to facilitate AI/ML application development and would extend, at the minimum, to include the variety of standard techniques currently in wide use, such as random forests (RF), multilayer perceptrons (MLP-ANN, a form of artificial neural network), and convolutional neural networks (CNN). Since platforms such as Google TensorFlow are already in wide use, it would be sensible to leverage those capabilities in developing this library.

A critical inclusion in such a data library should be reforecast datasets. These datasets, although computationally costly¹ to produce, are extraordinarily valuable for validating weather events, addressing calibration issues, and general predictability studies. Larger ensembles are valuable for providing proper baselines for probabilistic

¹ EMC estimates that a 30-yr global ensemble at 0.25° grid spacing, with 11 members run two times per week out to 48 days and 5 members daily out to 16 days, would require 1,500 million core-hours at a computational cost of approximately \$56 million. The primary cost in producing the reforecast is the reanalysis (B. Gross, V. Tallapragada, and J. Whitaker 2022, personal communication).

forecasting. The reforecast has the additional benefit of providing a stable dataset which will provide more robust weights in usual AI/ML tools such as ANNs, since the underlying data-generating mechanism (the computer model) is not changing. An example of the effective use of such datasets is the use of quantile mapping for precipitation forecasts within the National Blend of Models (NBM; Hamill et al. 2017; T. M. Hamill 2022, personal communication). However, owing to the cost of producing such ensembles, the number of members and the archived output has been restricted, making these datasets less useful. This expense should be supported, and available variables should be increased for the purpose of postprocessing in general and especially AI/ML work.

We note that “smart sub-sampling” (Kravtsov et al. 2022) is an approach that, within limits that depend on the application, can make reforecasts less costly. Those authors built a high-dimensional empirical model of temperature and precipitation that could produce a minimal subset of dates that provide representative sampling of local precipitation distributions across the contiguous United States, both in training and independent test data. To generate this model, however, a long time series of (reanalysis) data are needed.

Concomitant with reanalysis efforts should be the collation of relevant observations/analyses. An example of the latter is the need for long time series of quality high-resolution analyses in Alaska and Hawaii to improve the NBM. Further, convenient formatting of such datasets drastically improves efficiency of AI/ML/postprocessing efforts (e.g., chunked netCDF datasets for easy access and reduced data-wrangling time). The production of such benchmark datasets, organized according to agreed-upon standards and frameworks, would be a major step forward in facilitating AI/ML development efforts.

A key element of this concept is the need for AI/ML consultants who can facilitate the development and use of these tools by domain experts across the NWS. The challenge of finding the right balance between domain experts, craft consultants, and interdisciplinary agents is significant, but this team-based approach would allow these consultants to partner with NWS experts on specific projects of interest to those organizations without dispersing that expertise into the many existing silos. Additionally, this would allow for developing institutional knowledge concerning ongoing projects and reduce duplication of effort. This partnering will likely lead to the added, crucial development of in-house AI/ML expertise within those specific areas through the project basis of that activity.

Another example of how team-based approaches add value is the issue of feature/predictor selection. Time and attention employing meteorological intuition is necessary to determine potentially useful features, and to limit the constraints imposed by the “curse of dimensionality”—the size of needed training data increases exponentially as the dimensions of the AI/ML problem expand, so efforts to select and reduce relevant features are important.

Where this library is located within the NWS is immaterial to the overall concept and should be driven by logistical considerations—one likely location for it might be the MDL, given the extensive experience with postprocessing within that group. Notably, with the advent of the COVID pandemic and the success of virtual work across the NWS, it should be possible to establish a kind of hybrid organization for this library, which in the competitive environment for AI/ML expertise will allow for less difficulty in staffing.

This concept should work well for individual Weather Forecast Offices as well as the NWS Centers, provided that sufficient human effort is provided within the library for those individuals to work as collaborative development and implementation teams. This latter is obviously crucial as demand for such partnering is likely to be substantial, given the present and likely future activity in AI/ML within the NWS.

Fundamental AI/ML research. Currently, academic expertise is brought to bear through a variety of mechanisms including the JTTI; the Collaborative Science, Technology, and Applied Research (CSTAR) program; and Cooperative Institutes. Adjusting the beginning RL required for funding opportunities to sync more realistically with academic research would allow for the exploratory AI/ML work that is needed. Either some funding vehicle specific to that concept needs to be implemented or existing ones should be appropriately adjusted. The exploratory niche is clearly the most obvious place for academic work and fits well with NWS needs in the larger sense. For example, calibration works in opposition to the rare event nature of many of the forecasts of most interest—how to balance these considerations within a specific operational context necessitates coordinated, exploratory efforts along with input from operational experts.

Additionally, there is a need for a seamless, end-to-end pathway through which projects could pass from exploration/development to testbed to proving ground to operational implementation, along with the necessary personnel for ongoing support of those efforts. This formalized process must entrain NWS personnel from the beginning to retain crucial operational domain expertise. Theme-based calls for such efforts (e.g., heavy rainfall, wildfires, etc.), connected to a process as detailed above, would likely lead to more rapid progress than is currently possible. While there has been some effort to accomplish this in recent years through the JTTI and CSTAR programs, this has proven insufficient for the following reasons. First, these programs do not support the fundamental AI/ML research that is needed at the earliest stages, as discussed above in the context of RLs. Second, while these programs *can* lead to operational implementation, the path is not strictly end-to-end as described above. Third, the entrainment of NWS personnel, in practice, is limited owing to operational and other time constraints—in order for these relationships to be more effective, time must be set aside for that activity. Last, the number of awards that result from these programs is not sufficient to cover the breadth of the effort that is needed.

Future advances. Boukabara et al. (2021) argue that in NOAA, AI/ML will largely supplement, rather than replace, current tools and approaches. This follows the history of the use of guidance tools in operational forecasting. A new approach, the production of data-driven forecast models instead of computationally expensive numerical models (e.g., Weyn et al. 2020, 2021; Pathak et al. 2022; Lam et al. 2022), may be one exception. Employing a deep-learning model trained with reanalysis data, Weyn et al. (2021) were able to generate 85,800 reforecasts in a few hours on a single GPU. This model provides only a few output variables at 1.4° latitude–longitude grid spacing, lacks conservation laws or any direct incorporation of physics, and its skill is approximately a decade behind current forecast models at short-to-medium range. However, the model can learn physics-based phenomena directly from the data, and physical constraints such as conservation laws can be built into the learning process. Further, Weyn et al. (2020, 2021) show that it is straightforward to add additional variables to such models, and additional efforts of this kind have already shown skill competitive with NWP models (Pathak et al. 2022; Lam et al. 2022). Likely some combination of data-driven modeling will exist alongside of traditional NWP in the future.

The ability to rapidly generate large multimodel forecast ensembles that include initial condition uncertainty using the above approach can provide an additional operational benefit. It is sometimes the case that prior to a major weather event, details concerning the controlling factors are poorly known. One such example was the 3 May 1999 tornado outbreak in Oklahoma and Kansas (Roebber et al. 2002). As noted by those authors, prior to this event “no observational, conceptual, or NWP model evidence existed to support an outbreak scenario.” Their analysis, using “potential vorticity surgery,” indicated that the likelihood of an outbreak scenario was highly sensitive to details concerning an upper-level flow feature.

Another example is Ribeiro et al. (2022), who used a 40-member ensemble to demonstrate the low short-range predictability of a derecho event related to convection initiation, the organization of a dominant bow echo mesoscale convective system (MCS), and MCS maintenance. Such a capability does not exist in operations today, but could be feasible in the future using data-driven models.

Further, by using a cluster analysis tool, forecasters would be able to quickly identify the most likely outcome and the most likely worst-case outcome from the large ensemble of forecasts. A second application of cluster analysis to the initial conditions of these respective forecast scenarios might reveal particular elements that should be monitored most closely as the forecast evolves. This ability would improve situational awareness and operational forecast confidence. Such cluster approaches have already been applied to east coast winter storms (Zheng et al. 2019).

This essay has identified obstacles to achieving those advances but also offers solutions that will permit the great potential that AI/ML offers to operational weather forecasting by the NWS to be fully realized. We close with a statement from Boukabara et al. (2021), with which we wholly agree, and which we believe underlines the urgency to remove the obstacles to AI/ML work within the NWS:

Despite the challenges in leveraging AI for Earth science, we expect greatly expanding use of AI for environmental data and forecasting applications. The drive to simultaneously improve forecast skill (by accounting for unknown or difficult to model phenomena) and increase efficiency (therefore reducing cost and meeting latency requirements) will continue to make AI attractive to operational centers like NOAA.

Acknowledgments. Many individuals representing the following groups within NOAA contributed to the understanding described in this essay. These included the Aviation Weather Center, the Climate Prediction Center, the Environmental Modeling Center, the Meteorological Development Laboratory, the National Hurricane Center, the National Water Center, the National Environmental Satellite, Data, and Information Service, the NOAA Center for Artificial Intelligence, the Detroit Weather Forecast Office, the NWS Western Region, the Ocean Prediction Center, the Office of Science and Technology Integration, the Operational Proving Ground, the Physical Sciences Laboratory, the STI Modeling Program Team, the Space Weather Prediction Center, the Storm Prediction Center, and the Weather Prediction Center. Additional contributions were made by representatives of Colorado State University, Stony Brook University, the University of Oklahoma, the University of Washington, and the Climate Corporation.

Data availability statement. No datasets were generated or analyzed for this essay.

References

- Barrett, L. F., 2021: This is how your brain makes your mind. MIT Technology Review, 25 August, www.technologyreview.com/2021/08/25/1031432/what-is-mind-brain-body-connection/.
- Boukabara, S.-A., and Coauthors, 2021: Outlook for exploiting artificial intelligence in the Earth and environmental sciences. *Bull. Amer. Meteor. Soc.*, **102**, E1016–E1032, <https://doi.org/10.1175/BAMS-D-20-0031.1>.
- Cains, M. G., and Coauthors, 2023: Exploring what AI/ML guidance features NWS forecasters deem trustworthy. *22nd Conf. on Artificial Intelligence for Environmental Science*, Denver, CO, Amer. Meteor. Soc., 8A.2, <https://ams.confex.com/ams/103ANNUAL/meetingapp.cgi/Paper/419371>.
- Chase, R. J., D. R. Harrison, A. Burke, G. M. Lackmann, and A. McGovern, 2022: A machine learning tutorial for operational meteorology. Part I: Traditional machine learning. *Wea. Forecasting*, **37**, 1509–1529, <https://doi.org/10.1175/WAF-D-22-0070.1>.
- Dueben, P. D., M. G. Schultz, M. Chantry, D. J. Gagne II, D. M. Hall, and A. McGovern, 2022: Challenges and benchmark datasets for machine learning in the atmospheric sciences: Definition, status, and outlook. *Artif. Intell. Earth Syst.*, **1**, e210002, <https://doi.org/10.1175/AIES-D-21-0002.1>.
- Foster, K. R., and H. Kokko, 2008: The evolution of superstitious and superstition-like behaviour. *Proc. Roy. Soc.*, **276B**, 31–37, <https://doi.org/10.1098/rspb.2008.0981>.
- Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, [https://doi.org/10.1175/1520-0450\(1972\)011<1203:TUOMOS>2.0.CO;2](https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2).
- Hamill, T. M., 2015: New directions in statistical post-processing. *Harry R. Glahn Symp.*, Phoenix, AZ, Amer. Meteor. Soc., 47 pp., https://www.weather.gov/media/mdl/AMSGlahnSymp2015_Hamill_ppt.pdf.
- , E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, 2017: The U.S. national blend of models for statistical postprocessing of probability of precipitation and deterministic precipitation amount. *Mon. Wea. Rev.*, **145**, 3441–3463, <https://doi.org/10.1175/MWR-D-16-0331.1>.
- Haupt, S. E., W. Chapman, S. V. Adams, C. Kirkwood, J. S. Hosking, N. H. Robinson, S. Lerch, and A. C. Subramanian, 2021: Towards implementing artificial intelligence post-processing in weather and climate: Proposed actions from the Oxford 2019 workshop. *Philos. Trans. Roy. Soc.*, **A379**, 20200091, <https://doi.org/10.1098/rsta.2020.0091>.
- , and Coauthors, 2022: The history and practice of AI in the environmental sciences. *Bull. Amer. Meteor. Soc.*, **103**, E1351–E1370, <https://doi.org/10.1175/BAMS-D-20-0234.1>.
- Kodelja, Z., 2019: Is machine learning real learning? *Cent. Educ. Policy Stud. J.*, **9**, 11, <https://doi.org/10.26529/cepsj.709>.
- Kravtsov, S., P. Roebber, T. M. Hamill, and J. Brown, 2022: Objective methods for thinning the frequency of reforecasts while meeting post-processing and model validation needs. *Wea. Forecasting*, **37**, 727–748, <https://doi.org/10.1175/WAF-D-21-0162.1>.
- Lam, R., and Coauthors, 2022: GraphCast: Learning skillful medium-range global weather forecasting. arXiv, 2212.12794v1, <https://doi.org/10.48550/arXiv.2212.12794>.
- Lazo, J. K., H. R. Hosterman, J. M. Sprague-Hilderbrand, and J. E. Adkins, 2016: Impact-based decision support services and the socioeconomic impacts of winter storms. *Bull. Amer. Meteor. Soc.*, **97**, E626–E639, <https://doi.org/10.1175/BAMS-D-18-0153.1>.
- Litt, X., 2020: Xavier Litt: Chess shows that humans and AI work better together. *Irish Examiner*, 17 January, www.irishexaminer.com/opinion/commentanalysis/arid-30975938.html.
- McKenna, M., A. Boddy, and S. D. Baum, 2020: 2020 survey of artificial general intelligence projects for ethics, risk, and policy. Global Catastrophic Risk Institute Tech. Rep. 20-1, 156 pp., <https://gcrinstitute.org/2020-survey-of-artificial-general-intelligence-projects-for-ethics-risk-and-policy/>.
- NOAA Science Advisory Board, 2021: A report on priorities for weather research. NOAA Science Advisory Board Rep., 119 pp., https://sab.noaa.gov/wp-content/uploads/2022/04/PWR-Report-in-Brief_4March2022_Final.pdf.
- NSTC, 2019: The National Artificial Intelligence Research and Development strategic plan: 2019 update. NSTC PUBID-06-21-2019-001-01, 50 pp., www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf.
- Pathak, J., and Coauthors, 2022: FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. arXiv, 2202.11214v1, <https://doi.org/10.48550/arXiv.2202.11214>.
- Ribeiro, B., S. Weiss, and L. Bosart, 2022: An analysis of the 3 May 2020 low-predictability derecho using a convection-allowing MPAS ensemble. *Wea. Forecasting*, **37**, 219–239, <https://doi.org/10.1175/WAF-D-21-0092.1>.
- Roebber, P. J., 2022: A review of artificial intelligence and machine learning activity across the United States National Weather Service. NOAA Tech. Memo. NWS MDL 86, 25 pp., <https://vlab.noaa.gov/documents/6609493/8249989/TechMemo86.pdf>.
- , D. M. Schultz, and R. Romero, 2002: Synoptic regulation of the 3 May 1999 tornado outbreak. *Wea. Forecasting*, **17**, 399–429, [https://doi.org/10.1175/1520-0434\(2002\)017<0399:SROTMT>2.0.CO;2](https://doi.org/10.1175/1520-0434(2002)017<0399:SROTMT>2.0.CO;2).
- Schaffer, J. D., P. J. Roebber, and C. Evans, 2020: Development and evaluation of an evolutionary programming-based tropical cyclone intensity model. *Mon. Wea. Rev.*, **148**, 1951–1970, <https://doi.org/10.1175/MWR-D-19-0346.1>.
- Shermer, M., 2008: Patternicity: Finding meaningful patterns in meaningless noise. *Sci. Amer.*, **299** (6), <https://doi.org/10.1038/scientificamerican1208-48>.
- Stuart, N. A., and Coauthors, 2022: The evolving role of humans in weather prediction and communication. *Bull. Amer. Meteor. Soc.*, **103**, E1720–E1746, <https://doi.org/10.1175/BAMS-D-20-0326.1>.
- Weyn, J. A., D. R. Durran, and R. Caruana, 2020: Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002109, <https://doi.org/10.1029/2020MS002109>.
- , ———, ———, and N. Cresswell-Clay, 2021: Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002502, <https://doi.org/10.1029/2021MS002502>.
- Zheng, M., E. K. M. Chang, and B. A. Colle, 2019: Evaluating U.S. East Coast winter storms in a multimodel ensemble using EOF and clustering approaches. *Mon. Wea. Rev.*, **147**, 1967–1987, <https://doi.org/10.1175/MWR-D-18-0052.1>.