CLIMATE CHANGE AI DATASET WISHLIST

This is a list of datasets whose availability could accelerate progress or remove bottlenecks in different areas on the intersection of machine learning and climate change. We classify potential datasets according to a topic or area within our paper [1], and the following types of data availability:

1. Type 1: public, needs structure
2. Type 2: private, needs to be released
3. Type 3: scattered, needs collation
4. Type 4: proxies, needs to be inferred
5. Type 5: scarce, needs collection

1 Overarching datasets

1.1 Satellite imagery for remote sensing

Paper area: Electricity Systems, Buildings & Cities, and Farms & Forests

Description: Satellite (and/or aerial) imagery of various spatial and spectral resolutions with global coverage and at different (granular) time slices. There are numerous applications of this data in agriculture (soil composition, crop yield, and crop type detection), forest and high-risk ecosystem monitoring (tree height/type and land cover estimation), cities (building height estimation), energy (wind turbine and solar panel localization), and across sectors (methane, CO2, and N2O measurement). Different applications require different spatial, spectral, and temporal resolutions and different kinds of labels. For example, for many monitoring applications in the energy, building and transport sectors, spatial resolution is far more important and very high-resolution RGB images are often used. Spectral resolution is more relevant for instance for vegetation and land use.

ML challenge: Gathering information from image proxies (using a combination of satellite data and different sources of labeled data described in the rest of this document), and producing high-resolution images from multiple low-resolution images. This involves techniques from object detection, segmentation, and transfer learning.

Availability: Type 2, Type 3

Next Steps:

- High-resolution RGB images: Publicly available data can be aggregated from the US Geological Survey (see EarthExplorer), the Copernicus dataset (which provides free data from the Sentinel satellites), and NASA Worldview. Google Earth provides a cloud-free satellite image of the whole world, but this cannot necessarily be used for research without permission, since many of the images originally come from commercial providers. Commercial satellite images are for example available through DigitalGlobe (up to 31cm resolution) and Planet (up to 72cm resolution). These organizations have proven interest in providing images for research purposes, though these images may be expensive or otherwise difficult to access.
- Multispectral images: Providers such as DigitalGlobe, Planet, and Sentinel offers multispectral images (5-13 infrared bands) at the cost of spatial resolution. For example Sentinel-2 provides 13 spectral bands with 10 meter resolution. This spectral information can be highly valuable for making predictions related to biomass.
- Hyperspectral images: These images can contain up to a few hundred contiguous bands on the spectrum and can detect concentration of certain chemicals in the atmosphere such as GHGs. The increase in spectral resolution usually comes with a decrease in spatial resolution e.g. PRISMA contains 249 spectral bands with a 30 meter spatial resolution. See Table 1 of Survey of Hyperspectral Earth Observation.

		Availability Type	CCAI solution domains	Data Challenge	ML Challenge
1	Overarching Datasets				
	Satellite Imagery (remote sensing)	2,3	Electricity systems, Buildings & Cities, Farms & Forests	Availability & interactions between imagery of different spatial and temporal resolution**	Producing High-Res images from Low-Res imagery, Learning from multisensor data **
	LiDAR (topography)	1,2,3,4,5	Electricity systems, Buildings & Cities, Farms & Forests	Spatial coverage	Predict LiDAR for areas with limited coverage **
	Data for materials design	1,2,3	Electricity systems, CO2 removal, Industry	Materials dataset with physical properties	Characterize material for clean energy technologies
	Power grid systems	1,2,5	Forecasting Supply & Demand, Infrastructure	Multimodal sensor data	Demand response, Supply & Load forecasting, optimal power flow, economic dispatch problems, predictive maintenance
	Wind & Solar power prediction	1,2,3,4,5	Forecasting Supply & Demand	Spatial granularity	Power production forecasting (multiple time scales)
	Aggregated electricity demand		Forecasting Supply & Demand	Spatial & temporal granularity	Demand forecasting (multiple time scales)
2	Disruptive Future Technologies				
	Nuclear fusion data	2,3	Accelerating Fusion Science	Real-time sensor data with diverse temporal resolutions and experimental runs	Predicting disruptions
3	Battery Operation and Optimization				
	Battery operation & degradation	3	Alternative Fuels & Electrification, Transportation	Battery operations (voltage, capacity curves, degradation, usage)	Forecast & optimize battery operations
4	Transportation				
	Transportation arrival statistics	1,2,3	Modal shift	Multimodal (Bus, rail, freight)	Forecasting arrival times, spatial differentiation
	Freight activity	3,5	Transport (Reducing transport activity, Modal shift), Industry (Supply chains)	Freight activity data by transport mode	Origin-destination patterns
	Traffic counts	1,3,5	Reducing transport activity, Modal shift	Ground-based counts of vehicles by road, direction, vehicle class, geolocation	Mobility patterns
	Electric vehicle charging data	5	Alternative fuels & electrification	Charging & driving patterns	Predicting charging behavior, causal inference
	Urban mobility data	1(rare),2,3,4,5	Transport (Reducing transport activity, Modal shift), Building & cities (The future of cities)	Multimodal & geolocated passenger flow	Multimodal routing, origin-destination patterns, optimal dispatch/pricing
5	Building & Cities				
	Building level energy consumption	2,4,5	Optimizing buildings, Urban Planning	Multi-temporal individual building consumption (electricity & gas)	Demand prediction, pattern detection, grid optimization
	Building thermal dynamics		Optimizing buildings	Building temperatures including HVAC dynamics, occupancy	Optimal control of energy systems, modeling dynamical systems, grid flexibility services
	Building information	1,2,3,4,5	Urban Planning, The future of cities	Building stock (age, usage, number of floors, geometry)	Predicting urbanization patterns, travel demand, energy usage for heating and cooling, attributes and geometries
	City-level energy & GHG emissions	1,3,4,5	Building & Cities, Electricity systems, Transportation	City-level energy consumption & GHG emissions by sector, Fuel mixes, transportation data	Energy & GHG emission estimates, prioritizing mitigation actions
6	Industry				
	Product-specific GHG emissions	2,4,5	Supply chains	Life-cycle analyses database of common products & components	Suggest climate-friendly products/components
	Electrocatalysis data	2,5	Production & Energy	Database on electrocatalysis for GHG-intensive processes, including metadata on reactions	Designing novel electrocatalyts with lower activation temperatures (requiring less energy)
	Factory power for incrased energy efficiency & demand response	2,3,4,5	Production & Energy	Power usage data for industrial equipement & processes with associated GHG emisisons.	Optimize power-intensice processes, increase energy efficiency of industrial processes & HVAC with RL
	Methan leak data	2,3,4	Reducing current-system impacts	High resolution methane leakage data with relevant metadata	Detect methane leaks
7	Climate Mitigation				
	Geophysical data	2,3	Carbon dioxide removal	Data relevant for CO2 site identification	Predict complex outputs (e.g. seismograph reading, subsurface velocity CO2 migration)
8	Climate Prediction				
	Climate model outputs for fast approximation	1,2,3	Uniting data, ML & climate science	Aggregation of climate model outputs under varying conditions	TBD
9	Societal Impacts				
	Water treatment data	2	Infrastructure	Sensor & failure log data from water treatment plants	Predicitve maintenance (Optimal inspection schedule)
	Flood maps	3	Crisis	High spatial & temporal historical water-level data given flooding events	Flood forecasting & vulnerability assesment
	High resolution crop yield	3	Social systems	High spatial & temporal crop yield in diverse geographies	Forecast crio yield for food security interventions
	Social media posts for disaster relief	3	Crisis	Posts referring to Climate-related disasters	Post categorization to support crisis response, highlight actionable posts
	Forest fire	3	Managing forests	Real-time high resolution MWIR satellite data	Predict likelihood of forest fires
	Landslide occurences due to extreme weather events	1,2	Social systems, Crisis	Historical landslide data due to extreme weather with associated information on terrain, faults, geological xtics, landuse etc	Predict landslide occureance and Causal inferences between data features and landslides
10	Individual actions & Collective decisions				
	Consumer climate preferences	2	Facilitating behavior change	Data on consumer behavior & preferences for sustainable products	Recommender system for susistainable programs or technologies
	Global climate policy	1,3	Assessing & Informing policy	Dataset on climate change legislation	Evaluating transpareny, trust, equality

Figure 1: Summary of the CCAI dataset wishlist.

- Annotations: For monitoring of high-risk ecosystems, different conservation organizations generate annotations of ecosystem health, and they sometimes make it public online (e.g., ICIMOD, Chesapeake). There have already been some competitions in this area (e.g., deforestation on Kaggle) – they tend to partner with these organizations and make their data more accessible to ML specialists.
- Some applications may benefit from combining different sources of satellite images.

1.2 LiDAR data for topography

Paper area: Electricity Systems, Buildings & Cities, and Farms & Forests

Description: LiDAR data provides information on the topography of the Earth's surface. These data are particularly useful in mitigation applications involving buildings and in adaptation.

ML challenge: LiDAR is often used as ground truth for training on satellite images, which are more widely available. LiDAR is only available for limited spatial extent.

Availability: Type 1, Type 2, Type 3, Type 4, Type 5

Next Steps: Those data are available publicly or at a cost through various providers.

1.3 Data for materials design

Paper area: Electricity Systems, Industry, CO2 Removal

Description: Dataset of materials (e.g. their molecular/crystal structure) and their physical properties/performance.

ML challenge: To more quickly and efficiently discover, characterize, and classify materials for novel clean energy technologies (e.g. solar fuels, next-generation batteries, and next-generation solar cells), more efficient CO2 sorbents (e.g. for CO2 removal), and novel GHG-efficient replacements for carbon-intensive materials (e.g. steel, cement, and concrete).

Availability: Type 1, Type 2, Type 3

Next Steps: Aggregate, contextualize, and standardize data from publicly-available materials datasets, which may be difficult for ML practitioners to initially understand or parse.

2 Electricity systems

2.1 Power grid systems

Paper area: Forecasting supply and demand

Description: Electrical power grid is a network of subsystems each of which belongs to one of the following types: generators (supply), transmission lines, transformers, and loads (demand).

ML challenge: Supply and Load forecasting, demand response. Surrogate modeling of networked dynamical systems across wide spatial and temporal scales. Optimal power flow, economic dispatch problem.

Availability: Type 1, Type 2, Type 5

Next Steps: Aggregate, contextualize, and standardize data from publicly available materials datasets.

2.2 Solar power production for forecasting

Paper area: Forecasting supply and demand

Description: A dataset with solar power production amount at a spatially granular scale (e.g. for every generator, small group of generators, or farm) and at a temporally granular scale (e.g. every five minutes or every hour). This data should be time- and location-stamped, and accompanied by information/features including weather measurements and (potentially) images/videos of the sky.

ML challenge: To forecast solar power production at different time scales (e.g. 5 minutes, 1 hour, a few hours, days, months, or years ahead) and at different spatial scales (e.g. generator-level, state-level, system operator-level). These forecasts should be accurate and characterize uncertainty. They should potentially be explainable and/or account for system goals (e.g. forecasts might be judged based on how well they inform low-emissions electricity system operation, rather than being judged solely on accuracy).

Availability: Type 1, Type 2, Type 3, Type 4, Type 5

Next Steps:

- Power production and location data (large-scale generators): Can be collected from solar operators or system operators who likely have this data.
- Power production and location data (small-scale generators): Likely involves a mix of collecting small-scale solar panel size/location data from government agencies and augmenting these (likely incomplete) datasets using ML methods to detect panel size/location. Power production can then be computed using physical models that use these size/location data along with weather data.
- Weather and sky image/video data: Can likely be aggregated from government agencies and (high-resolution) satellite imagery providers.

2.3 Wind power production for forecasting

Paper area: Forecasting Supply & Demand

Description: A dataset with wind power production amount at a spatially granular scale (e.g. for every generator, small group of generators, or farm) and at a temporally granular scale (e.g. every five minutes or every hour). This data should be time- and location-stamped, and accompanied by information/features including weather measurements.

ML challenge: Similar to that in "solar power production," except for wind instead of solar.

Availability: Same as in "solar power production," except that sky image/video data is likely not needed.

Next Steps: Same as in "solar power production," with the following differences:

- Power production and location data (large-scale generators): In the United States, this data can also be collected from government agencies, as they collect some of this information to give out wind production tax credit (PTC) incentives.
- Sky image/video data is likely not needed (but weather data likely is).

2.4 Aggregated electricity demand for forecasting

Paper area: Forecasting Supply & Demand

Description: A dataset with electricity demand aggregated to spatial and temporal levels that do not expose individual consumer habits. (This level should be decided upon in consultation with power system operators and government agencies to protect consumer privacy.) This data should be time- and location-stamped, and accompanied by information/features including weather measurements.

ML challenge: Similar to that in "Granular solar power production," except for demand instead of solar, and at potentially less granular spatial/temporal scales depending on data availability.

Availability: Type 1, Type 2, Type 3

Next Steps:

- Demand data: In certain parts of the United States, aggregated demand data may be publicly available through Independent System Operators or Regional Transmission Operators (ISOs/RTOs) and can simply be pulled from their websites. In the case where more granular data is needed, or in regions (within and outside the US) where such data is not public, it may be necessary to coordinate with relevant system operators.
- Weather data: Can likely be aggregated from government agencies.

2.5 Power grid data for predictive maintenance

Paper area: Infrastructure

Description: Records describing breakdowns and maintenance on a real power grid. This will be multimodal, and at minimum should include: Sensor measurements of system ("feeder") failures, Time series of power loads, linkable to sensor failures, Logs documenting failure events (These "tickets" may give more context into the type of failure), Network graph structure

ML challenge: Given a prespecified maintenance budget, describe an optimal inspection schedule.

Availability: Type 2

Next Steps: These data are held privately by energy companies. The template for this is work by Cynthia Rudin with ConEdison – her collaborators on that project may be able to inform the curation of such a dataset.

3 Disruptive future technologies

3.1 Nuclear fusion data for disruption detection

Paper area: Accelerating fusion science

Description: Real-time sensor data from tokamak fusion reactors at various temporal resolutions that is captured for a large number of experimental runs ("shots"). Each shot either ends in a disruption, or is non-disruptive; this information should also be included.

ML challenge: Predicting the onset of disruptions with sufficient warning time. (Disruptions are instabilities with the potential to damage the fusion machine in question and delay further fusion experiments.)

Availability: Type 2, Type 3

Next Steps: The data for any given tokamak is private to the management organization of that device. Access usually requires a "collaboration agreement" and explicit authorization from the organization in question. Given this landscape, the next steps likely include selecting the tokamaks for the dataset, and then contacting the associated management organizations. Organizations/machines include EUROfusion (JET tokamak), General Atomics (DIII-D tokamak), etc. The management organizations may not be willing to release this data publicly (as the machine funders paid for the data and thus may not have incentive to give it away), so it may be necessary to come up with creative arrangements.

4 Battery operation and optimization

4.1 Battery operation and degradation datasets

Paper area: Alternative fuels and electrification

Description: A dataset of battery operation including current, voltage characteristics, capacity curves, etc. Includes both operation cases and degradation information. Dataset for major commercial battery technologies.

ML challenge: Forecast and optimize battery operation, adapt batteries to various environments and use cases (various high impact papers in the area recently).

Availability: Type 3

Next Steps: These datasets are available in various publications, so some degree of standardization is required in addition to collection efforts.

5 Transportation

5.1 Transportation arrival statistics for infrastructure design

Paper area: Modal shift

Description: For forecasting arrival times to improve the speed and reliability of public and multimodal transportation. This includes passenger bus and rail, and freight rail. Larger datasets for all transportation modes and many different geographical areas are needed, as currently those are typically available on a case study basis (for specific transportation modes and cities).

ML challenge: Forecasting arrival times, understanding how models generalize between different geographies (e.g. public transit systems in different cities).

Availability: Type 1, 2, and 3

Next Steps: Cities and countries might make data available for certain geographic locations. The first goal would be to have a collection of cleaned datasets that allow both to improve forecasting on the specific task, and experiment with generalization across different geographic locations.

5.2 Freight activity for optimizing transport

Paper area: Transport (Reducing transport activity, Modal shift), Industry (Supply chains)

Description: These data consist of freight activity (weight carried x distance in e.g. ton-km) by transportation mode at various levels of granularity. This can for example involve data that show subnational commodity flows with origin-destination pairs and details about the commodity, but also data that compare the flows of goods and modal shares globally. Freight activity is not surveyed and reported in many countries, especially for road and water freight.

ML challenge: Analysis of origin-destination patterns, such as routing efficiency or demand forecasting

Availability: Type 3, Type 5

Next Steps: Countries and states might have data available for certain geographic locations (see for example the US Commodity Flow Survey). Ocean vessels are tracked by vessel traffic services through the automatic identification system, which can for example be accessed through the U.S. Coast Guard.

5.3 Traffic counts for infrastructure design

Paper area: Reducing transport activity, Modal shift

Description: Traffic monitoring is done by counting the vehicles that pass a road with ground-based counters. Typically, counts are reported as counts per hour, travel direction, vehicle class, and geolocation of the counting device. Some jurisdictions make those counts publicly available.

ML challenge: ML can provide information about mobility patterns – which is directly necessary for agent-based travel demand models, one of the main transport planning tools. For example, ML makes it possible to estimate origin-destination demand from traffic counts.

Availability: Type 1, Type 3, Type 5

Next Steps: Since many states and countries publish this data, a first step would be to compile a list of sources and information about their quality. Such datasets include: UTD19 is a large-scale traffic data set from over 20000 stationary detectors on urban roads in 40 cities worldwide making it the largest multi-city traffic data set publically available. See https://utd19.ethz.ch/ for more information.

5.4 Electric vehicle charging data for infrastructure design

Paper area: Alternative fuels and electrification

Description: Sensors in electric vehicles that record charging and driving patterns. Potential privacy issues.

ML challenge: Predict aggregate charging behavior of electric vehicle owners, causal inference

Availability: Type 5

Next Steps: In-vehicle sensors might become increasingly available to record this type of data. Those datasets need to be collected, anonymized and made public through organizations.

5.5 Urban mobility data for infrastructure design

Paper area: The future of cities, Transportation (Reducing transport activity, Modal shift)

Description: Passenger flows geolocated in real-time across urban transportation mode (cars, buses, trains, scooters, bikes, pedestrians)

ML challenge: Map-matching, trip start/end detection, clustering and analysis of origin-destination patterns, optimal dispatch/pricing, multimodal routing

Availability: Type 1 (rarely), Type 2, Type 3, Type 4, Type 5

Next Steps: These data are often proprietary and scattered. In order to understand how to reallocate flows for one mode to another, data on as many modes as possible would need to be integrated, but each operator owns its data. Few cities mandate all the data from vehicle sharing operators to be public. There are technical and privacy issues, in particular with pedestrian data. A possible next step is platforms at the city scale and operated by a dedicated organism, with either data donations from users, or data from companies if there is a collaboration with, or a mandate from the city.

6 Buildings and Cities

6.1 Building level energy consumption for demand management

Paper area: Optimizing buildings, Urban planning

Description: Aggregated (electricity, gas) consumption at the individual building scale and at different levels of temporal granularity, from hourly to yearly.

ML challenge: Demand variability prediction, demand prediction at scale, clustering patterns, grid optimization

Availability: Type 2, Type 4, Type 5

Next Steps: Consultation with system operators. In certain regions, system operators own this data (e.g. Europe/US). When the operator is public, like in France, the context may be different. An important question is how much data is available in developing countries.

6.2 Building thermal dynamics for optimal energy management

Paper area: Optimizing buildings

Description: Thermal dynamics modeling evolution of the building temperatures, including HVAC dynamics with actuator signals, and disturbance signals such as ambient temperature, solar irradiation, occupancy data. Datasets at the individual building scale and at different levels of temporal granularity, sampling time from minutes to hours, datasets length from days to years.

ML challenge: Modeling dynamical systems, Optimal control of energy systems, grid flexibility services

Availability: Type 1, Type 2, Type 3

Next Steps: These data are often proprietary and scattered. Besides aggregating openly available datasets we need need to convince several data owners to share their data publicly.

6.3 Building information for urban planning

Paper area: Urban planning, The future of cities

Description: Maps of building stock that include place, geometry and energy-relevant attributes (usage, approximate year of construction, number of floors, etc.).

ML challenge: Prediction of attributes and geometries, Prediction of energy use for heating and cooling, analysis/prediction of urbanization patterns, analysis of destinations and attraction points in cities (e.g. for travel demand prediction)

Availability: Type 1, Type 2, Type 3, Type 4, Type 5

Next Steps: The most comprehensive dataset is Google Maps, but the data are proprietary. OpenStreetMap proposes a free alternative, but quality and completeness varies across cities. There is further data at the city level (e.g. cadaster) that need further integration. To fill the gaps, citizen science (people mapping on the street) and inference with ML can be used. A next step can be to support current data infrastructures like OpenStreetMap to help them populate faster energy-relevant categories, as climate change is not their primary focus.

6.4 Urban mobility data for infrastructure design

Paper area: The future of cities, Transportation (Reducing transport activity, Modal shift)

Description: Passenger flows geolocated in real-time across urban transportation mode (cars, buses, trains, scooters, bikes, pedestrians)

ML challenge: Map-matching, trip start/end detection, clustering and analysis of origin-destination patterns, optimal dispatch/pricing, multimodal routing

Availability: Type 1 (rarely), Type 2, Type 3, Type 4, Type 5

Next Steps: These data are often proprietary and scattered. In order to understand how to reallocate flows for one mode to another, data on as many modes as possible would need to be integrated, but each operator owns its data. Few cities mandate all the data from vehicle sharing operators to be public. There are technical and privacy issues, in particular with pedestrian data. A possible next step is platforms at the city scale and operated by a dedicated organism, with either data donations from users, or data from companies if there is a collaboration with, or a mandate from the city.

6.5 City-level energy and GHG-emissions data

Paper area: Buildings & Cities, Electricity Systems, Transportation

Description: City-level energy consumption by sector, GHG emissions by sector, transportation activity, fuel mixes, etc.

ML challenge: Energy or GHG emission estimates, inter-city comparison, prioritizing mitigation actions

Availability: Type 1, Type 3, Type 4, Type 5

Next Steps: Different organizations have aggregated this data for limited sets of cities and with varying indicators. Indicators need to be standardized to create a consistent dataset across as many cities as possible.

7 Industry

7.1 Product-specific GHG emissions for informed consumption

Paper area: Supply chains

Description: Both consumers and manufacturers cannot choose climate-friendly products unless they can compare the embedded GHG emissions of various alternatives. A database of life-cycle analyses of popular components and products (combined with transportation options) could provide information about GHG emissions embedded in every production process from mining/extraction through manufacture and distribution.

ML challenge: To offer suggestions for which potential vendors, components, and/or products in a supply chain are most climate-friendly for any given location, and to extrapolate about other types of products that lack GHG emissions data.

Availability: Type 2, Type 4, Type 5

Next Steps: The challenge here is to provide enough life-cycle data to be able to make useful suggestions, but the data that exist in this space are incredibly scattered and often proprietary or unmeasured. Many climate-conscious consumers and brands do prefer to purchase sustainable products, although the lack of information in this space allows for substantial "greenwashing" based on marketing rather than data. A carbon tax would make many of these hidden supply chain emissions more visible and mandate better reporting, while providing a strong price signal for consumers. In the meantime, larger and more vertically-integrated firms could start publishing GHG information about GHG embedded in their manufacturing and supply chain, and requiring their suppliers to follow suit.

7.2 Electrocatalysis data for lowering the temperature of industrial processes

Paper area: Production and energy

Description: A comprehensive database of electrocatalysts for ammonia and similar GHG-intensive processes, along with relevant information about the reactions themselves to design effective training sets.

ML challenge: Ammonia generation and other industrial-scale chemical processes rely exclusively upon fossil fuels for heating up reactants, and ML could contribute to designing novel electrocatalysts that could lower the activation temperatures required for reactions to occur.

Availability: Type 2, Type 5

Next Steps: ML researchers must collaborate with chemical engineering labs and industrial manufacturers to create databases of machine-readable data for designing industrial catalysts.

7.3 Factory power for increasing energy efficiency and demand response

Paper area: Production and energy

Description: Industry and/or factory-specific data on the power usage of industrial equipment and processes, ideally combined with information about expected GHG emissions of electrical power given the weather and time of day (related to the dataset for Aggregated Electricity Demand above). Datasets on electricity conversion, energy storage, and reconversion pathways that use surplus electric power, typically during periods where fluctuating renewable energy generation exceeds load.

ML challenge: Increase energy efficiency of industrial processes and heating, ventilation, and air conditioning (HVAC) systems through reinforcement learning. Optimize power-intensive processes to run off renewable energy, or at least minimize the need for GHG-intensive peak power sources. Model complex pathways in industrial power to X systems. Optimize energy efficiency of the industrial processes. Constrained optimal control. Safety of the operation.

Availability: Type 2, Type 3, Type 4, Type 5

Next Steps: Convince companies to share their electrical data, and coordinate with the providers of industrial automation equipment (Rockwell Automation, Siemens, ABB, Honeywell, etc) to facilitate effective data collection and analysis across their customer firms. Also, corporate datasets must be combined with electrical utility data to ensure that factories are minimizing GHG emissions in addition to electricity cost.

7.4 Methane leak data for detection/prevention

Paper area: Reducing life-cycle fossil fuel emissions

Description: Dataset with methane leak locations and magnitudes at natural gas pipelines, compressor stations, power plants, etc. These should be accompanied by features including sensor data, and ideally by high-resolution satellite data.

ML challenge: Detect methane leaks, either preemptively or retroactively, so natural gas companies or other entities can fix them.

Availability: Type 2, Type 3, Type 4

Next Steps: Sensor reading data likely needs to be aggregated from natural gas companies, who privately hold this data. (These companies may or may not have leak detection data, but likely have an incentive not to release it.) The SLED project led by the Southwest Research Institute (SwRI) and US Department of Energy's National Energy Technology Laboratory (NETL) may have some sensor data and data about previously-detected leaks that they may be willing to share. The company Bluefield Technologies is deploying microsatellites with high resolution hyperspectral cameras meant to detect methane leaks, and could potentially be approached for data.

8 Climate mitigation

8.1 Geophysical data for CO2 sequestration site suitability / subsurface identification

Paper area: CO2 removal

Description: Gather datasets useful for identifying promising sites for CO2 sequestration, or for monitoring those sites once they become active.

ML challenge: Mapping from limited data to complex outputs (e.g. seismograph readings, subsurface velocity models, or monitoring well data, and CO2 migration status)

Availability: Type 2 and/or Type 3

Next Steps: Oil and gas companies likely have a lot of real-world experimental data (with "ground" truth) that could be directly helpful. There are also many works that use data from expensive simulations.

9 Climate Prediction

9.1 Climate model outputs for fast approximation

Paper area: Modeling (also relevant for geoengineering/control)

Description: This dataset would aggregate the outputs of many, many climate model runs under various conditions, in some nicely structured form.

ML challenge: Ideally, there would be many possible ML problems one could try to extract from this dataset, at scales from the subgrid to the whole planet.

Availability: Type 1, Type 2, and Type 3 probably all exist.

Next Steps: Need to decide on which GCM to emulate at what resolution, ideally in consultation with many other research groups. Then there needs to be agreement on data format for consistent aggregation. Climate model emulators (e.g., FaIR, MAGICC) emulate different general circulation models (GCMs) for different emission scenarios.

10 Societal Impacts

10.1 Water treatment data for predictive maintenance

Paper area: Infrastructure

Description: A combination of sensor and failure log data from an operating water treatment plant.

ML challenge: Given a prespecified maintenance budget, describe an optimal inspection schedule.

Availability: Type 2

Next Steps: These data are collected by municipal water treatment plants. Different city departments of water may be open to curating and sharing these data, and there are many research papers about optimizing wastewater treatment plants, but the data generally do not seem to be public.

10.2 Flood maps for infrastructure design and urban planning

Paper area: Disaster relief

Description: Historical water-level data associated with flooding events, with fine spatial and temporal resolution.

ML challenge: These data could be applied to two ML challenges: (1) forecast floods, to provide effective alerts and (2) assess infrastructure vulnerability.

Availability: Type 3

Next Steps: There appear to be public sources for many regions, though they seem somewhat inaccessible, and it's unclear how granular the measurements are.

10.3 High-resolution crop yield for food security

Paper area: Food security

Description: Average crop yield at (better-than) county level spatial granularity. Ideally, several crops would be measured simultaneously, yields would be available over time, and distinct geographies would be represented. These yields would be linked temporally and spatially with satellite imagery and meteorological data.

ML challenge: From satellite imagery and meteorological data, forecast crop yield at a timescale at which appropriate food security interventions could be made.

Availability: Type 3

Next Steps: These types of data are often collected by governments or curated by agriculture experts. Yield data are sometimes publicly available online but aren't curated for machine learning, and are sometimes woefully limited in granularity.

10.4 Social media posts for disaster relief

Paper area: Disaster relief

Description: Social media posts that refer somehow to specific climate-related disasters – floods, wildfires, and strong storms are ideal candidates. A few types of disasters, affecting different parts of the world, should be represented.

ML challenge: Organize posts in a way that supports crisis responders: cluster related posts and highlight immediately actionable ones.

Availability: Type 3

Next Steps: A first step is to collect tweets with hashtags related to a few representative disasters of interest. It may be worthwhile to request expert annotation categorizing tweets into types of action items. This public dataset would emulate the private data that this company curates – they may be able to offer suggestions.

10.5 Forest fire

Paper area: Earth monitoring systems, active detection and alerts

Description: Ability to triangulate soil and climate data to predict and show at risk areas of forest fires. non-classified access to real-time high-resolution MWIR satellite data to quickly detect fires when they are small enough to put out easily.

ML challenge: Predict the likelihood of forest fires based on spatiotemporal data.

Availability: Type 3

Next Steps: Collect and aggregate openly available data from public sources. Process raw data into ML-readable formats.

10.6 Occurrence of landslides to detect new landslides due to extreme weather events

Paper area: Disaster relief, Risk and resilience

Description: Inventory of past landslides that occurred due to rainfall thresholds exceeding their maximum limit. The database ideally contains information that helps to explain the causes of the landslide: presence of faults, terrain relief, geological composition, land use (forest, grass, urban), accumulated rainfall and soil moisture, and predict landslides before they occur.

ML challenge: Prediction and causal inferences based on various geological, meteorological, land use, and other features. Support alert systems in detecting the occurrence of landslides early based on past landslide data.

Availability: Type 1, Type 2

Next Steps: These types of data are usually collected in a disaggregated manner by governments. But they are not aggregated in a way that is conducive to machine learning and are sometimes limited in granularity.

11 Individual actions and Collective decisions

11.1 Consumer climate preferences for facilitating behavior change

Paper area: Modeling consumer behavior and facilitating behavior change

Description: A dataset which could allow for the assessment of which consumers are most amenable to sustainable behavior change, or products such as smart home devices that could reduce energy consumption. May include user demographics or previous purchasing history, especially as related to sustainable products. Additionally, could include consumer's stated preferences on climate-related questions, such as how much they would be willing to pay to subsidize sustainable programs that reduce energy consumption or environmental harm.

ML challenge: To cluster consumers together based on their preferences, and identify groups of consumers that are most interested in new sustainability programs or new technologies that could lead to reduced consumption. A recommender-system approach could also be taken here.

Availability: Type 2

Next Steps: Identify a set of climate-relevant products and services that could be used to distill a given consumer's climate preferences Determine how to release this data publicly. Ideally, collect additional data about consumer's explicit climate preferences.

11.2 Global climate policy dataset

Paper area: Assessing policy options

Description: A comprehensive global dataset of legislation passed to address climate change. This can help track progress and pinpoint where policy change is needed. Ideally, the dataset would include full text files and metadata about instrument types, sectors addressed, etc. ML challenge: To cluster current policy impact and look at the gaps

ML challenge: Transparency, trust, equality.

Availability: Type1, Type 3

Next Steps: Some databases already exist that track climate policies, which is usually based on a large manual collection and description effort.

References

[1] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, A. Luccioni, T. Maharaj, E. D. Sherwin, S. K. Mukkavilli, K. P. Körding, C. P. Gomes, A. Y. Ng, D. Hassabis, J. C. Platt, F. Creutzig, J. T. Chayes, and Y. Bengio, "Tackling climate change with machine learning," *CoRR*, vol. abs/1906.05433, 2019. [Online]. Available: http://arxiv.org/abs/1906.05433