2020 Building Performance Analysis Conference and
SimBuild co-organized by ASHRAE and IBPSA-USA

TOWARDS A STANDARD CLIMATE DATA MODEL FOR BUILDING DESIGN AND ANALYSIS

Sagar Rao[1] and Parag Rastogi[2]
[1]NeuMod Labs, Madison, WI, USA
[2]arbnco, Glasgow, Scotland, UK

ABSTRACT

The design and analysis of buildings includes many aspects such as resilience, energy, water, occupant experience, and operations and maintenance. In relation to climate analyses alone, specialized techniques are being developed to study past climate, leverage real-time weather information, plan for forecasted climate, tackle microclimate effects, and the like. Some advanced automation workflows employ non-traditional simulation engines using targeted reduced order modeling techniques.

Although these workflows are evolving at a rapid pace, climate data models have not kept up with them. When simulation tools were developed, many implemented dissimilar climate information models, resulting in competing formats. So, while engine-neutral data is sometimes represented using mutually incompatible data structures and data dictionaries, other types of climate information may be unavailable because these files are restricted to inputs supported by their respective simulation engines. Also, the separate distribution of statistical data such as design day information, climate zones classification, etc., is a frequent source of design and analysis inconsistency.

This paper demonstrates why a standard definition is necessary for climate information to implement intelligent automation workflows and to take advantage of contemporary computing techniques.

INTRODUCTION

There are many components to climate information required for building design and analysis, but they fall into three broad categories – i) design conditions (historic extremes), ii) time-series data for physical quantities (measured and synthetic), and iii) statistically-derived information for climate-driven decision making.

Currently, this data is available from multiple sources and in many different file types. Most of these data dictionaries, schemas, and file formats were developed at different times and for different simulation engines, so they now look somewhat ad hoc and do not lend themselves well to web data exchange, especially at the data layer in engine-neutral software architectures. Most of these files are distributed in delimited plain text format, and the most popular of these are EnergyPlus Weather (EPW) files (NREL & USDOE, 2017).

Although the plain text format can be space efficient, files generated by different data vendors frequently include varying degrees of detail, resulting in inconsistent scripting inputs. These plain text files also do not employ an extensible key-value pair data structure so it is a high-risk endeavor to develop holistic modeling frameworks and workflows, especially when end-uses for climate information are evolving rapidly, and many of these go beyond traditional energy simulation.

Therfore, as energy modelers try to achieve greater workflow automation for design consulting, it is imperative that a web-first data model be developed to represent richer climate information using contemporary data structures that are implemented in a modern file format.

This paper presents details of current sources for raw climate information; describes existing and evolving use cases for climate data; and explains why it is necessary to develop a standard data model to represent this information. It also makes recommendations for what would constitute such a data model and briefly discusses file formats for implementation.

CLIMATE INFORMATION

The words ‘weather’ and ‘climate’ are often incorrectly used interchangeably in building design and analysis. “Climate is the synthesis of weather events over the whole of a period statistically long enough to establish its statistical ensemble properties (mean value, variation, probabilities of extreme events, etc.) and is largely independent of any instantaneous events” (Essenwanger, 2001). To design infrastructure for a location that is fit for purpose during its lifetime, e.g., 25-30 years for an HVAC system, we need information about the climate – the so-called ‘design extremes’. For analysis, on the other hand, we need weather time series that are representative of the climate, i.e., samples that represent

the range of conditions that are seen in a climate. The typical meteorological year files are an example of that – composed of weather time series from 'representative' months for a given climate. Thus, a climate data model must include a means to represent weather time series that can be connected to other weather time series through a geographical identifier to create a representation of climate.

Climate Data Sources

Historical weather information for the entire world is available free of charge from the Integrated Surface Database (ISD, NCEI, 2020). The database is composed of time series from individual weather stations, i.e., they represent measurements from a fixed physical point (except sea-based and temporary stations). The data from the ISD alone is usually not enough to create a weather 'file' for analysis because very few stations record solar data. So, a user would need to supplement this source with data purchased from national meteorological organizations (many organizations do not share their complete datasets with the ISD), e.g., national solar radiation database (NSRDB) (Wilcox, 2012), and satellite consortia, e.g., Copernicus (European Commission, 2020). Satellite data are available for 'cells', i.e., trapezoidal areas of the earth defined by their latitudinal and longitudinal extent. Private and institutional providers have made available curated versions of this data specifically for building analysis (e.g., Huang, 2012, 2015), building design (Ch. 14, ASHRAE, 2017).

Time series of projected changes in climate, the so-called 'Climate Change Models' can be accessed for free through the CORDEX collaboration (WRCP, 2017). Each time series is provided by a national/regional meteorological agency and several agencies may cover the same location. The time series is generated from a Regional Climate Model (RCM), a physics-based model of climate for a region that is seeded/bounded by a Global Climate Model (GCM). Regional models cover overlapping domains, e.g., West and South Asia, Europe. The outputs are organized into cells of approximately 50 km on end (0.44^o x 0.44^o).

Finally, a relatively new source of weather time series is reanalysis datasets, e.g., NASA's MERRA-2 (Gelaro et al., 2017). These are global/regional climate models, like those mentioned before but instead of future projections, these models provide uniform, complete time series for estimated past weather conditions over a uniform global grid. These are calibrated to measured data where available and can serve as useful substitutes for areas with worse data availability. New work from ASHRAE is examining the feasibility of using reanalysis data in place of measured data to expand availability of design conditions and analysis data (RP-1745, ASHRAE, 2020).

As with any database, the data quality and reliability are independent of the format. An EPW file may be composed of low-quality data or high-quality. The quality of weather data, as with any other kind of data, depends on sources, processing algorithms, and fitness for purpose. For example, a Typical Meteorological Year (TMY) is a composite (artificial) year of weather data for a given location made up of months selected from many years of historical record using a specific algorithm (Wilcox & Marion, 2008). It may be available as an EPW file or as a table in Excel – in both cases it remains a TMY. The TMY file from a given provider must use valid data sources and the correct algorithm. A TMY file may also be composed of a sufficient sample of plausible future weather time series, e.g., those created using a stochastic weather generator (Eames et al., 2011; Rastogi, 2016; Lowe et al., 2018).

Data Distribution Formats

So far, we have discussed data *sources*. The data from these sources are stored in a variety of file formats and structures. For example, CORDEX data is available as netCDF4/HDF5 files (Unidata, 2020), while the ISD stores data in fixed-width flat files. The data are not always at the resolution required for simulation, though design conditions can be calculated from synoptic (summary) data. The variety of data sources and formats discussed here points to the need for an extensible, modular climate data model able to account for the needs of storing time series with varying metadata.

Providers of simulation-ready files stitch together data from different sources and formats into suitable hourly data structures. In some cases, interpolation and regression are required to up-sample 3- or 6-hourly readings to hourly readings, or to infer missing data (Rastogi, 2016, and references therein). A unified data exchange model would, therefore, need to maintain compatibility with both the formats used by those collecting or producing the original data, e.g., NCEI, NASA, and those using it, e.g., EnergyPlus users. The advantage of using a key-value structure is that it can be used for applications such as machine-learning, automation, data-based consulting in addition to simulation with a specific engine.

Several providers distribute historic and synthetic weather files as EnergyPlus weather (EPW) files. EPW is a delimited plain text file format with a well-defined data dictionary, but a non-standard file format. Therefore, their content, layout, meta-data, and data may vary depending on the source generating them. Commented lines can contain memos, but these are unstructured and inconsistent across providers and

sources. Other commonly used file formats for weather data include binary files using the DOE-2 data dictionary (DOE-2 Reference Manual Part 1, version 2.1, 1981, p. 2), custom comma separated values (CSV) files, weather files for ESP-r (Clarke, 2001), and WEA, the weather file for DAYSIM/Radiance (Ward, 1994). For each one of these, the lack of defined key-value pairs always requires a data dictionary lookup as these files are not comprehensible on their own, and are seldom human-readable.

CURRENT USE CASES

Climate Data for Design

Climate data is regularly used in building design for sizing, distribution, and installation of building systems. Typical data consists of extremes for solar radiation, dry bulb temperature, web bulb temperature, dew point temperature, wind speed, wind direction, precipitation, etc. The primary objective of these calculations is to derive a "peak load" for various operational conditions to ensure sufficient system capacity to meet the project's design goals (ASHRAE, 2017). Similarly, designers may use precipitation data for the sizing of catchments or water distribution systems; and solar radiation data for preliminary sizing of solar arrays, or to specify envelopes that meet the minimum performance criteria from an energy code such as ASHRAE Standard 90.1 (ASHRAE et al., 2019).

Most design teams start with developing a Basis of Design document (BOD) that establishes appropriate design conditions to use for sizing calculations such as – thermal load analyses, coil sizing, and cooling tower sizing, etc.. They may then obtain these inputs from the ASHRAE Handbook of Fundaments (HOF, ASHRAE, 2017). The HOF makes this data available in a table for project documentation. This design information is also regularly used by energy modelers to inform equipment sizing. Engineers and analysts routinely work together so energy analysis can inform equipment sizing and optimize control sequences for complex systems. For instance, EnergyPlus uses input macro files (IMF) and design day files (DDY) for sizing and control. DDY files format the design conditions from the handbook into "design days" that can be incorporated into EnergyPlus. EnergyPlus EPWs are also accompanied by STAT files. The STAT files are particularly helpful for design as they contain statistical values such as Universal Thermal Comfort Index, Monthly Average RH, Climate Zones, etc. However, the downside of having three separate files is that they need to be distributed together. Conversely, although they may be obtained together by a designer, there is no way of verifying that they are based on the same dataset, thus resulting in inconsistent inputs for design and analysis. DOE-2 (DOE-2 Reference Manual Part 1, version 2.1, 1981) users include these values in their input (INP) files, or manually specify them using an application like eQUEST (doe2.com). eQUEST design days fall within one of two categories – heating or cooling. Both design days define similar physical properties. Similar files are also available for other simulation engines, such as Trane TRACE 700 (trane.com).

Climate Data for Analysis

Weather data for simulation is generally formatted in a table of hourly values for meteorological quantities and generally consists of 8760 (365days x 24hrs) entries, though finer resolutions are permitted by many formats. It can be one of three types: i) historic (measured), ii) typical (synthetic), or iii) future (synthetic). A historic weather file consists of a full year of measured data formatted to work with a specific simulation engine. The other two types are classified as 'synthetic' since they are both artificially composed. A typical year is usually composed from many years of weather records, though the individual months are selected from a given year, each based on how closely their distribution matches the overall distribution for that month. Thus, unlike a historic weather file, a typical weather file does not represent an actual year (Crawley & Huang, 1997; Rastogi, 2016, sec. 2.6.1), but rather a representation of an artificial 'median year'. Finally, future files are purely synthetic, even if they are transformed versions of measured time series using techniques such as morphing.

Though several 'typical year' algorithms have been proposed over the years (Crawley & Huang, 1997), they share the common theme of compressing many years of weather information representing a climate into a single composite year. Various providers use different source years, e.g., the TMY15 (climate.onebuilding.org) uses the last available 15 years of data from the year of publication. When using these datasets, however, the accompanying design conditions must also be selected from the same dataset to maintain consistency between design and performance analysis, i.e., design conditions based on the same record period as the one used to derive the TMY. This would not matter in a stable climate, i.e., one where the statistical properties that define a climate, such as – monthly means and variance are constant over time. However, climate change makes the inconsistences more apparent, especially in view of the rapid and significant changes in recent decades. Finally, calculating design conditions from a typical file itself is meaningless since a typical file deliberately excludes extremes. The presence of an hourly value that is close to a design condition in any typical file is purely

coincidental. For some analyses this can be a limitation, and alternate selection algorithms have been proposed, e.g., eXtreme Meteorological Year (XMY) (Crawley & Lawrie, 2019).

DEVELOPING USE CASES

As high-performance buildings start to expand their performance goals beyond energy, we find that data to calculate other aspects of building design such as climate resilience, water efficiency, occupant experience does not always exist in legacy data models. For example, the lack of a standard or reliable source of rainfall data requires water modelers to currently obtain and parse these inconsistent data sets on their own. The use of standard naming conventions and modern data mining techniques could help significantly reduce the time needed to parse these datasets for use in water modeling and compliance reporting. We now discuss only some of the emerging use cases for climate information in building design and performance evaluation and why a corresponding evolution in the existing climate data models is necessary to accommodate them.

Future Climate & Resilience

As the effects of climate change become more pronounced, dedicated climate consulting is gaining prominence, e.g., to help firms comply with TCFD reporting requirements (TCFD, 2019) or assess climate-based risks more broadly.

The so-called "future weather files" are developed using one of two techniques – morphing (Belcher et al., 2005), and stochastic generation (e.g., Rastogi, 2016). Morphing acts directly on the 'seed' time series, i.e., measured data for a station, to scale and shift it based on climate model predictions. Stochastic generation instead scales and shifts the historical distribution of parameters after suitable transformations for seasonality and non-normality (Boland, 1995). The time series generated are formatted into single-year files and written out in an engine-specific file format (EPW, BIN, etc.) such that the forecasted hourly values can be used for hourly or sub-hourly simulations without any changes to the engines. When these techniques produce future typical monthly profiles (like the TMYs), the so-called FTMY (Future Typical Meteorological Year), they do so by either morphing/transforming an existing typical file or generating a number of samples for a future time period and then applying a selection algorithm like TMY. Since the FTMYs also represent median years, they do not include extreme data deliberately.

Future time series generation techniques create a unique problem for climate data models since they can be used to create samples (single weather time series) from the distribution of weather parameter values that represent a climate as it is projected to be at a future time. Each sample is a plausible future weather time series, but not an exact prediction of a specific time in the future. So while each is valid, none is definitive. Using projected time series therefore requires a climate data format that can accommodate uncertainty by identifying each sample output by a method (variant) uniquely. The results of simulating with several sample future time series can be collated to create an ensemble prediction of the performance of a building or system at some point in the future.

Dashboards that provide historic, present, and future context and provide data for UV Index, air quality, weather alerts, etc. are also being developed. Some of these physical quantities are not supported by the climate data models discussed here. This is primarily because most of them were developed for annual (and sub-annual) building simulations of thermal properties and lighting/daylighting.

The authors have previously proposed (Rastogi, 2016), and are currently developing, analysis methods that require or favor multi-year probabilistic modeling over single-year deterministic modeling. While energy modelers have been using programs such as the OpenStudio Parametric Analysis Tool (PAT) for parametric simulations for some time now, large-scale stochastic modeling of buildings is still used sparingly in building design. Probabilistic modelling differs from parametric simulations in several ways. Parametric modelling is usually used to choose the specific value of a design parameter systematically from a finite selection of possible values from a 'catalogue' of possibilities. Probabilistic modelling, on the other hand, would be used to determine the behavior of a system under random boundary/input conditions, sampled from a realistic distribution of these inputs. This is particularly suited to use cases where inputs are unknown or unknowable, e.g., design that is resilient or robust to climate change. As greater computing power becomes available and storage costs decrease, probabilistic modeling of multi-year climate files is a well-suited method for future climate consulting for the built environment. Multi-year simulations are conceptually simple but work best when done programmatically. Therefore, to make the most of these contemporary analysis techniques requires a streamlined climate information model that is well-defined, web-first, scalable, and extensible. In addition, as more projects undertake resilience/robustness analyses, climate studies that quantify the impact of hail, flooding, hurricanes, heat waves, cold waves, and other natural hazards are gaining popularity. Current weather files are inadequate to represent such climate information as they were not originally designed to record and distribute extreme weather events.

Climate Database for Design and Analysis

Several web-based modeling tools and workflows are now available that move away from traditional desktop-based computation of building performance to analyses carried out through remotely-deployed simulation engines and distributed computing, e.g., BuildSim (buildsim.io), Ladybug Tools (ladybug.tools), LightStanza (lightstanza.com). Although most of these tools replicate the conventional desktop user experience, their underlying architecture is fundamentally different. They consist of four key components: i) a web client to interact with models through a graphical user interface (GUI), ii) a server that provides computation and storage capabilities, and iii) a standard data model for storage, retrieval, and distribution of information to and from the central repositories, and iv) software development kits (SDKs) and application programming interfaces (APIs) that allow users to interact with the software platform for custom development.

In addition to this evolution in delivery technologies, there is a push for greater intelligent automation. Programming-based workflows are being developed for start-to-end automation that rely on consistency and reliability in input data for analysis, visualization, and reporting (Rao et al., 2018). These workflows rely heavily on standardization of data inputs and key-value type data structures lend themselves much better to this sort of intelligent automation. Climate data is a critical input to most analyses, therefore, a central database for climate information that can be accessed on demand using a standard data model would ensure consistency and repeatability across use cases and users, and significantly reduce the cost of developing simulation, visualization, and reporting scripts using the modern web infrastructure.

CURRENT DEFICIENCIES

We discuss some issues stemming from the very nature of legacy data models such as – i) the lack of standard naming convention for parameters and variables, ii) the existence of multiple generation methods for synthetic files that do not follow a regular update cycle and a way to track this information in the file itself, iii) the packaging of relevant information in multiple files, and iv) the use of legacy file formats that do not lend themselves well to clean-up and automation. Therefore, improperly-curated or inconsistently-handled climatic data is a quality control issue for individuals and teams making decisions over several stages of design and analysis.

Moreover, weather files distributed using legacy data models do not contain the same set of climate variables, and lack a common naming convention for them. This is primarily because these data dictionaries and structures were developed to represent weather data for specific simulation engines, so the representation of climate information is specific to what is required as inputs to each calculation method.

Climate data vendors collect, curate, and generate time series and statistical summaries for the various use cases mentioned here. While big institutional vendors develop their own data exchange formats, e.g., NetCDF, data providers specializing in building design and analysis, e.g., METEONORM (meteonorm.com) and White Box Technologies (weather.whiteboxtechnologies.com), tend to provide files in formats compatible with popular building simulation engines.

Further, regardless of whether the data is made available for free or for a fee, the publishers require copyright, fair-use conditions, and protection of IP. For example, posting a file purchased from a vendor on a public discussion board is usually a violation of their terms of fair use. This issue cannot be effectively policed without control over the distribution of human-readable files. This leads to a situation where a vendor prices their offering to account for the 'leakage' and unauthorized reuse, which drives up costs for all users, which promotes unauthorized reuse and the incentive to use obsolete free data sources. An accessible climate database that, instead, allows a user to access reports and simulation results without accessing the complete underlying dataset reduces this problem. This database could still be free, in part or wholly, but it would allow for consistent distribution of the best available and up-to-date climatic information.

DATA MODEL RECOMMENDATIONS

Climate Data vs Climate Data Model

In this section we introduce a proposed new data model for climate-based design and analysis of engineering systems. It is important to distinguish the *data model* from the *data* itself. We are not proposing a new set or type of climate data to replace the ones discussed before. Rather, we are proposing a new data model to enumerate, store, and exchange climate data across different use cases and calculation engines.

The key requirements for the next generation of climate information data model for Building Performance Modeling (BPM) are identified as follows:

- a single format for historic, typical, and future weather data,
- combined packaging of geographic/location information, design conditions, extreme climate, and weather measurements,
- the ability to expand beyond traditional tabular climate information models,

- support for multi-year probabilistic climate data,
- built-in intellectual property management to encourage the regular update and consistent distribution of high-quality climate data,
- consistent versioning to support a universal climate API,
- a well-maintained validator to facilitate the development of automation workflows using this data,
- mapping tools to and from existing file formats such as bin, epw, ddy, clm, wea, and
- a web-first approach to file format selection.

Data Structure

We propose a data structure consisting of key-value pairs and sorted lists of values as it lends itself well to web data exchange. This structure is also supported by most data interchange formats (e.g., objects and arrays in JSON), and has direct mapping to common programming language objects (e.g., dictionaries and lists in python). The proposed structure is also consistent with another data model standardization underway through ASHRAE Standard 205 – Standard Representation of Simulation Data for HVAC&R and Other Facility Equipment [in-review].

Like Standard 205, the climate data model can consist of a high-level collection of related items called Data Groups, which in turn consist of atomic data items called Data Elements. Each data group component can be assigned an Upper Camel Case (UCC) name, e.g. ClimateZone, while each data element, can be assigned a Lower Camel Case (LCC) name, e.g. standardPressure.

Model ID, Station Lookup & Metadata

Each climate file would have a universally unique identifier (UUID). A UUID is a 128-bit number which is practically unique if generated using a standard method (ISO/IEC 9834-8, 2014). Providing a unique ID will help track datasets, allow version control, help programmers write automation scripts, and facilitate restricted access to files, if necessary.

An accurate and reliable system of station identifiers is a foundational requirement for any climate data model. Traditionally, this has happened through station identifier schemes such as World Meteorological Organization (WMO) identifiers, International Civil Aviation Organization (ICAO) Location Indicators, Federal Aviation Administration (FAA) Location Identifiers, Weather Bureau Army Navy (WBAN) identifier, Master Station Library (MASLIB) Catalog Number, and others. However, most of these identifier schemes are mutually incompatible and some are poorly documented and published, which poses significant challenges to data exchange. Stations within some of these identifier schemes also get moved, resulting in further inconsistency. Therefore, here again, it is proposed that UUIDs be used to represent each station, tied to a unique latitude and longitude pair. So, a station that moves receives a new ID while one that is closed retains the ID that was assigned to it. Tying stations to their geolocation will allow better geotagging of stations and support greater integration with modern mapping and Geographic Information System (GIS) apps and APIs, such as ArcGIS (arcgis.com).

Moreover, time series generators such as climate data models and reanalysis outputs are specifically designed to output time series for a cell, an area on the globe defined by latitudinal and longitudinal extents. A cell may contain none to any number of weather stations. A consistent mapping or lookup capability between time series from synthetic weather generators and measured data, therefore, requires geotagging of the latter.

An extension of the geographical metadata fields proposed is rich metadata such as street address, elevation, climate zones, universal thermal comfort index (UTCI) etc. to ensure compatibility with existing tools and methods, and to enable automated lookup for these inputs. A simple example would be the inclusion of climate zones to assign envelope constructions to automatically create an ASHRAE 90.1 baseline for compliance with code and rating systems.

Design Conditions

Among others, two factors are key to how design conditions are handled in the data model – i) ensuring that they are generated and distributed with the "measured" values (described in next sub-section) to avoid inconsistencies in design documentation and analysis, and ii) shifting from "design days" to support for the full set of design conditions so a universal API can be developed to obtain any statistical design condition that is currently available in the HOF. "Design days" may be derived from these design conditions.

The following data groups represent high-level organization of design conditions:

- DesignConditionsAnnual,
- DesignConditionsExtremeAnnual,
- DesignConditionsMonthly,
- DesignConditionsMonthlyAverage.

This is consistent with the nomenclature defined in the HOF (Ch. 14, ASHRAE, 2017). Moreover, each of these data groups can then contain the individual design conditions that are listed in Table 1A of Chapter 14 (excerpt in Figure 1).

Table 1A Nomenclature for Tables of Climatic Design Conditions

CDD*n*	Cooling degree-days base n°C, °C-day
CDH*n*	Cooling degree-hours base n°C, °C-hour
DB	Dry-bulb temperature, °C
DBAvg	Average daily dry-bulb temperature, °C
DBSD	Standard deviation of average daily dry-bulb temperature, °C
DP	Dew-point temperature, °C
Ebn,noon	Clear-sky beam normal irradiances at solar noon, W/m²
Edh,noon	Clear-sky diffuse horizontal irradiance at solar noon, W/m²
Elev	Elevation, m
Enth	Enthalpy, kJ/kg base 0°C and 101.325 kPa pressure
HDD*n*	Heating degree-days base, n°C, °C-day
HR	Humidity ratio, $g_{moisture}/kg_{dry\ air}$
Lat	Latitude, °N
Long	Longitude, °E
MCDB	Mean coincident dry-bulb temperature, °C
MCDBR	Mean coincident dry-bulb temp. range, °C
MCWB	Mean coincident wet-bulb temperature, °C
MCWBR	Mean coincident wet-bulb temp. range, °C
MCWS	Mean coincident wind speed, m/s
MDBR	Mean dry-bulb temp. range, °C
PCWD	Prevailing coincident wind direction, ° (0 = North; 90 = East)

Figure 1: Excerpt from Table 1A Chapter 14 ASHRAE HOF

Measured Data

The time-series data for physical quantities can be assigned to a data group, "Measurements". A data element such as "quantities" can list the specific physical quantities, or a subset, of the enumerated measurements supported by the data model. For example, Dry Bulb Temperature, Wet Bulb Temperature, Dew Point Temperature, Relative Humidity, Direct Normal Irradiation, Diffuse Horizontal Irradiation, Wind Speed, Wind Direction, Precipitation. The enumerators can be developed to ensure full support for existing data dictionaries such as EPW, BIN, and WEA. The quantities list can be an unordered list of valid enumerators and function as column headers for subsequent data elements, each of which is a time-stamped measurement.

The Measurements data group can be expanded to store typical calculated quantities such as – solar time, solar altitude angle, solar azimuth angle, total irradiation on a tilted surface, etc. This would be particularly useful for reduced order engines that do not include sky models for building simulation and the creation of automated reports as they would negate the need to create single zone models.

The timestamped arrays can be extended to include multi-year support in a single file. Extremely large files can quickly become unwieldly to manage and transmit, and the same is true of existing formats, where several files would be needed to represent multiple years. However, the modularity and consistency of the proposed model reduces the need for transmission of complete files. Cloud-based services could access data as needed without exchanging complete files or tying up resources on a user's computer. For visualization and reporting APIs, only specific columns of information need be transmitted.

Data Licensing

Data groups such as Vendor, License, and Licensee must be considered to allow easier licensing and tracing of climate information. As mentioned earlier, new methods are being developed to generate typical and future climate files so native support for data licensing will enable greater market adoption from vendors generating this data. It will also help ensure that periodic updates can be made to a climate dataset with controlled access for certain datasets, if a data vendor chooses to do so.

IMPLEMENTATION FORMAT

The data model itself is independent of the file format. However, an implementation using a modern file format is required to demonstrate data exchange capabilities, and the building of robust APIs and SDKs. Although many file formats are available, the two most popular are eXtensible Markup Language (XML) and JavaScript Object Notation (JSON, Standard ECMA-404: The JSON Data Interchange Syntax, 2017). Both are widely used as data exchange formats and have also been adopted by applications to store structured data. The JSON file format is recommended for the first implementation of this data model due to its compatibility with objects in programming tools and popularity.

A strong argument in favor of JSON is its direct mapping to programming language objects through its key-value pairs, which makes JSON significantly easier to work with as a developer. JSON coding is comparatively simpler and code maintenance is also easier, thus reducing development time. Given that is a derivative of JavaScript, it is also the data interchange format of choice for most JavaScript plugins used for UI development, server management, data visualization, etc. This is also true for BPM tools that are being consciously developed for the web, e.g., epJSON inputs for EnergyPlus, OpenStudio Workflow (OSW) files for OpenStudio, and others.

A key drawback with JSON, however, is that it lacks support for logically-complex schemas. Although this would be a lesser concern for the climate data model, it could potentially affect future scalability of related data models and it makes it more difficult to build test suites and schema validators.

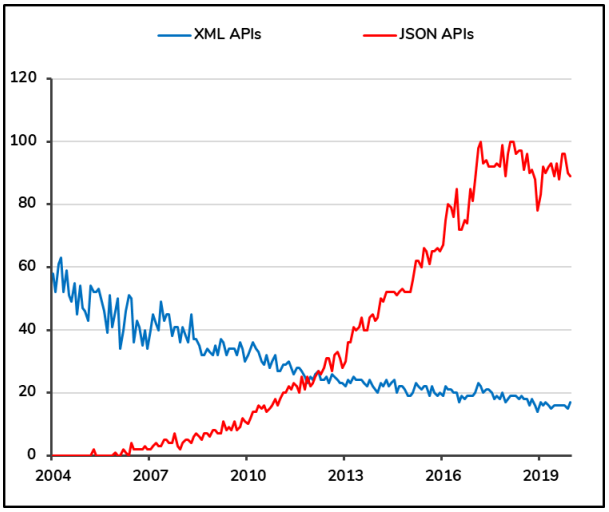

Figure 2: Google Trends Chart for XML API and JSON API searches worldwide between 2004 and 2019.

The authors recognize that although JSON is the current preferred format for data exchange for the web, when newer technologies become available, they may supplant JSON as the format of choice. Therefore, the development of the data model is disconnected from the file format to ensure compatibility with future software.

CONCLUSION

This paper provides a foundation for the development of a standard climate data model for building design and analysis. It identifies limitations in current data dictionaries, schema, and file formats to provide recommendations for the key components of a more contemporary climate data model for building design and analysis.

REFERENCES

ASHRAE. (2017). *2017 ASHRAE handbook.* (M. S. Owen, Ed.; Fundamentals). American Society of Heating Refrigerating and Air-Conditioning Engineers.

ASHRAE. (2020, March 27). *Meetings | ASHRAE 4.2 Climatic Information*.

ASHRAE, ANSI, & IESNA. (2019). *Standard 90.1-2019: Energy Standard for Buildings Except Low-Rise Residential Buildings* (Standard 90.1-2010; ASHRAE Standards).

Belcher, S. E., Hacker, J. N., & Powell, D. S. (2005). Constructing design weather data for future climates. *Building Services Engineering Research and Technology*, *26*(1), 49–61.

Boland, J. (1995). Time-series analysis of climatic variables. *Solar Energy*, *55*(5), 377–388. https://doi.org/10.1016/0038-092X(95)00059-Z

Clarke, J. A. (2001). *Energy Simulation in Building Design* (2nd ed.). Butterworth-Heinemann.

Crawley, D. B., & Huang, Y. J. (1997). Does it matter which weather data you use in energy simulations. *User News*, *18*(1), 25–31.

Crawley, D. B., & Lawrie, L. K. (2019). Should We Be Using Just 'Typical'Weather Data in Building Performance Simulation? Building Simulation 2019, Rome, Italy.

Eames, M., Kershaw, T., & Coley, D. (2011). On the creation of future probabilistic design weather years from UKCP09. *Building Services Engineering Research and Technology*, *32*(2), 127–142.

Standard ECMA-404: The JSON Data Interchange Syntax, (2017). https://www.json.org/json-en.html

Essenwanger, O. M. (2001). *General Climatology 1C: Classification of Climates* (Vol. 1C). Elsevier Science Limited.

European Commission. (2020, March 27). *Copernicus*. Copernicus: Europe's Eyes on Earth. https://www.copernicus.eu/en

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., … Zhao, B. (2017). The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454.

Huang, Y. J. (2012). *International Weather for Energy Calculations* (ASHRAE Research Project RP-1477; Development of 3012 Typical Year Weather Files for International Locations). ASHRAE. https://www.ashrae.org/resources--publications/bookstore/international-weather-for-energy-calculations

ISO/IEC 9834-8:2014, (2014). https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/27/62795.html

Lowe, J. A., Bernie, D., Bett, P., Bricheno, L., Brown, S., Calvert, D., Clark, R., Eagle, K., Edwards, T., Fosser, G., & others. (2018). UKCP18 science overview report. *Met Office Hadley Centre: Exeter, UK*.

NCEI. (2020, March 27). *Integrated Surface Database (ISD)*. National Centers for Environmental Information (NCEI)Formerly Known as National Climatic Data Center (NCDC).

NREL, & USDOE. (2017, November). *EnergyPlus*. https://energyplus.net/weather

Rao, S., Conant-Gilles, D., Jia, Y., & Carl, B. (2018). *Rapid Modeling Of Large And Complex High Performance Buildings Using Energyplus*. SimBuild 2018. Chicago, IL, USA.

Rastogi, P. (2016). *On the sensitivity of buildings to climate: The interaction of weather and building envelopes in determining future building energy consumption* [PhD, EPFL].

TCFD. (2019). *Task Force on Climate-related Financial Disclosures: Status Report*. TCFD.

Unidata. (2020, March 27). *NetCDF*. UCAR Community Programs. https://www.unidata.ucar.edu/software/netcdf/

Ward, G. J. (1994). The RADIANCE lighting simulation and rendering system. *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '94*, 459–472.

Wilcox, S. (2012). *National Solar Radiation Database 1991–2010 Update: User's Manual* (NREL/TP-5500-54824). NREL, National Renewable Energy Laboratory.

Wilcox, S., & Marion, W. (2008). *Users' Manual for TMY3 Data Sets*. http://www.nrel.gov/docs/fy08osti/43156.pdf

WRCP. (2017, April 1). *CORDEX*. World Climate Research Programme CORDEX. http://www.cordex.org/

York, D. A., Tucker, E. F., & Capiello, C. C. (Eds.). (1981). *DOE-2 Reference Manual Part 1, version 2.1*. Los Alamos Scientific Laboratory; Lawrence Berkeley Laboratory.