



OPEN

DATA DESCRIPTOR

CaRDS - the statewide California Residential water Demand and Supply open dataset


Marie-Philine Gross^{1,2} , Alvar Escriva-Bou³, Erik Porse⁴ & Andrea Cominola^{1,2} 

As water scarcity becomes the new norm in the Western United States, states such as California have increased their efforts to improve water resilience. Achieving water security under climate change, population growth, and urbanization requires an integrated multi-sectoral approach, where adaptation strategies combine supply and demand management interventions. Yet, most studies consider supply-side and demand-side management strategies separately. Water conservation efforts are mainly driven by policy requirements and publicly available data to assess the effectiveness of demand- and supply-side management policies is often hard to find and unstructured. Here we present CaRDS - the statewide California Residential water Demand and Supply open dataset. CaRDS encompasses nine years (2013-2021) of monthly water supply and demand time series for 404 water suppliers in California, USA, compiled from different open-access data sources. Access to detailed temporal and spatial water supply operations and demands at the state-level can be useful to researchers and practitioners to realize applications such as evaluating the effectiveness of water conservation policies and discovering regional differences in water resilience measures.

Background & Summary

In recent decades, droughts have become increasingly frequent and prolonged in the Western United States¹. Climate change is expected to exacerbate the severity of these extreme hydroclimatic events, putting more stress on the already scarce water resources in the region^{2,3} and increasing the vulnerability of water systems⁴. National and regional governments worldwide are implementing new multi-sectoral adaptation strategies and policies to adapt to climate variability and change in the pursuit of water security⁵. In response to mid-twentieth century water scarcity, the federal, state, and local governments in California built large-scale infrastructure to move water over long distances for use by cities and farms⁶⁻⁸. As new sources dwindled, the drought in 1976-77 instigated drought restrictions, which began several decades of state and federal policies to reduce water use through building codes and efficiency standards, especially for indoor use⁹. To counter the more severe droughts in recent times and in the future, the California State Legislature enacted two policy bills in 2018 to establish a new foundation for long-term water conservation and drought adaption planning - "Making Water Conservation a California Way of Life"^{10,11}. Water retailers will need to implement water demand management and conservation strategies to meet goals depending on local characteristics of efficiency investments, landscape irrigation, and land use characteristics¹².

Water suppliers have addressed water scarcity challenges through a mix of supply and demand-side measures. Water supply strategies focus on increasing water availability by developing new sources (e.g., surface supplies, groundwater, water reuse, and desalination) or by optimizing the management of existing sources. Water demand-side management focuses more on the consumers and their water use. Active areas of research are among others water demand forecasting, water consumption behavior change programs (e.g., via feedback on end-use water use activities or fixture upgrades), alternative water pricing schemes, and leak management. Existing literature generally considers demand-side management and supply-side measures separately, whereas

¹Chair of Smart Water Networks, Technische Universität Berlin, Berlin, 10623, Germany. ²Einstein Center Digital Future, Berlin, 10117, Germany. ³Civil and Environmental Engineering and Institute of the Environment and Sustainability, University of California, Los Angeles, CA, 90095-1496, USA. ⁴California Institute for Water Resources, University of California Agriculture and Natural Resources, Davis, CA, 95618, USA. e-mail: marie-philine.gross@tu-berlin.de

the concept of integrated water management recommends shifting from an isolated view on certain parts of water management and the water cycle to a more holistic approach.

A chance to realize such a multi-sectoral management strategy lies within the availability of FAIR (findable, accessible, interoperable, and reusable) data¹³. The deployment of digital technology and sensors in the water sector sparked the interest of researchers and practitioners towards an unprecedented amount of highly disaggregated data in time and space and related data driven analytics for both water supply and demand modelling and optimization¹⁴. A recent study showed that the digitalization of water utilities is globally underway, but its progress is highly dependent on individual management decisions¹⁵. This means big water data is often only available for more progressive water utilities and cannot be shared with the public due to security concerns. On the other hand, government organizations try to increasingly make public data more accessible by releasing data online for everyone to use. Most of the time, however, this data is not directly usable for analysis purposes¹⁶. Available data often lacks documentation, has inconsistent temporal resolution, or has machine-readability issues. This then requires extensive data processing, hampering a direct use of accessible data sources. For the reasons above, a large-scale, California-wide evaluation of the trade-offs of water supply and demand management strategies to support large-scale planning efforts for demand management in combination with supply-side interventions is not readily available.

Here we present CaRDS - the statewide *California Residential water Demand and Supply* open dataset, which contains monthly values of water supply and residential water demand disaggregated at the supplier level for 404 water suppliers in California from 2013 to 2021. CaRDS advances the literature in three ways: (i) data usability - this dataset closes the gap between data availability and the direct usability of public water supply and demand data sets; (ii) data consistency - CaRDS integrates continuous time series of monthly records for water supply and demand, together with corresponding data on precipitation and drought conditions for each supplier location. CaRDS thus gathers data that are currently scattered in various sources and reported with different formats, time/space aggregation levels, and units, making them consistent; (iii) data coverage - the dataset provides a consistent multi-annual statewide coverage of water suppliers in California, allowing research and applications that go beyond individual suppliers/regions and short-term analysis. CaRDS can be used to study the trade-offs between water supply and residential demand at the state scale, as well as to support regional studies, e.g., at the county, climatic zone, or hydrologic region scale. Further, water suppliers can compare their management strategies with neighboring suppliers. CaRDS can easily be extended with new data each year and it is possible to combine it with other data sets, e.g., for electricity or wastewater.

CaRDS is compiled based on three different public data sources: the Electronic Annual Reports (eAR) from the California State Water Resources Control Board¹⁷, the PRISM Climate Data from the PRISM Climate Group¹⁸, and the Climate Division Data (ClimDiv) from National Oceanic and Atmospheric Administration¹⁹. First, we extract water supplier information, the corresponding monthly water supply and residential demand from the eARs, and perform pre-processing steps to improve data consistency and enable data integration. We then use this information to geographically match the water suppliers with different climatic variables from PRISM and ClimDiv (see Fig. 1). In the following sections we describe CaRDS, the data processing steps, and its technical validation in more detail. Finally, we demonstrate how we can make open-access data workable.

Methods

CaRDS merges data from three different open-access sources to provide a temporally and spatially detailed description of water supply and residential water demand in California at the supplier level, along with hydro-climatic information for each location. In the following sections, we will first introduce the original data sources separately and explain the processing steps we performed for each type of data. We will then describe the data integration of the different data sources into one comprehensive dataset, CaRDS.

Electronic Annual Report. At the core of CaRDS are the water demand and supply data, which are extracted from the eAR. The eAR is a mandatory annual survey conducted by the State Water Resources Control Board (SWB) of California. As early as the 1980's, urban water agencies in California submitted detailed monthly water use data to the California Department of Water Resources. According to the California Water Law, the correct term to refer to such urban water agencies partaking in the data reporting would be *community water systems*. Here we will utilize the more commonly used term *water suppliers*. Over the course of decades, the frequency and detail of reporting varied. The SWB moved to an online survey in 2009, giving it today's name eAR. Starting in 2013, suppliers were required to submit detailed monthly data as part of drought management requirements, and the regular collection of detailed operations data was standardized as part of duties of the SWB. In 2019 the SWB released online the completed eARs starting from 2013. Here we use the online available surveys from 2013 to 2021¹⁷, in total nine years, to obtain information about water supply and demand at the supplier level for the whole state of California.

The reports encompass 3306 coded variables in the most current version, impeding a direct extraction of variables of interest. Additionally, the data quality is impacted by human error as data are manually reported by each water suppliers. This results in missing values, fluctuating reporting rates (despite the mandatory nature of the reporting), and an inconsistent encoding of string characters, which makes machine readability challenging. These data quality issues have greatly limited the past use of this data outside of the SWB.

For each year, we extract the following seven variables: supplier ID, monthly water production (supply), measurement unit of the water supply, total and residential monthly water deliveries (demand), measurement unit of water demand, and number of people served. Suppliers often also report agricultural, commercial, and institutional water demands, which are not included in CaRDS, as it focuses on residential water demand only. One of the challenges in data cleaning is that variable name conventions were changed by the SWB starting with the 2020 survey. Therefore, we process the survey data in two groups separately, one including the entries reported between 2013 and 2019, the second comprising those reported between 2020 and 2021. Another

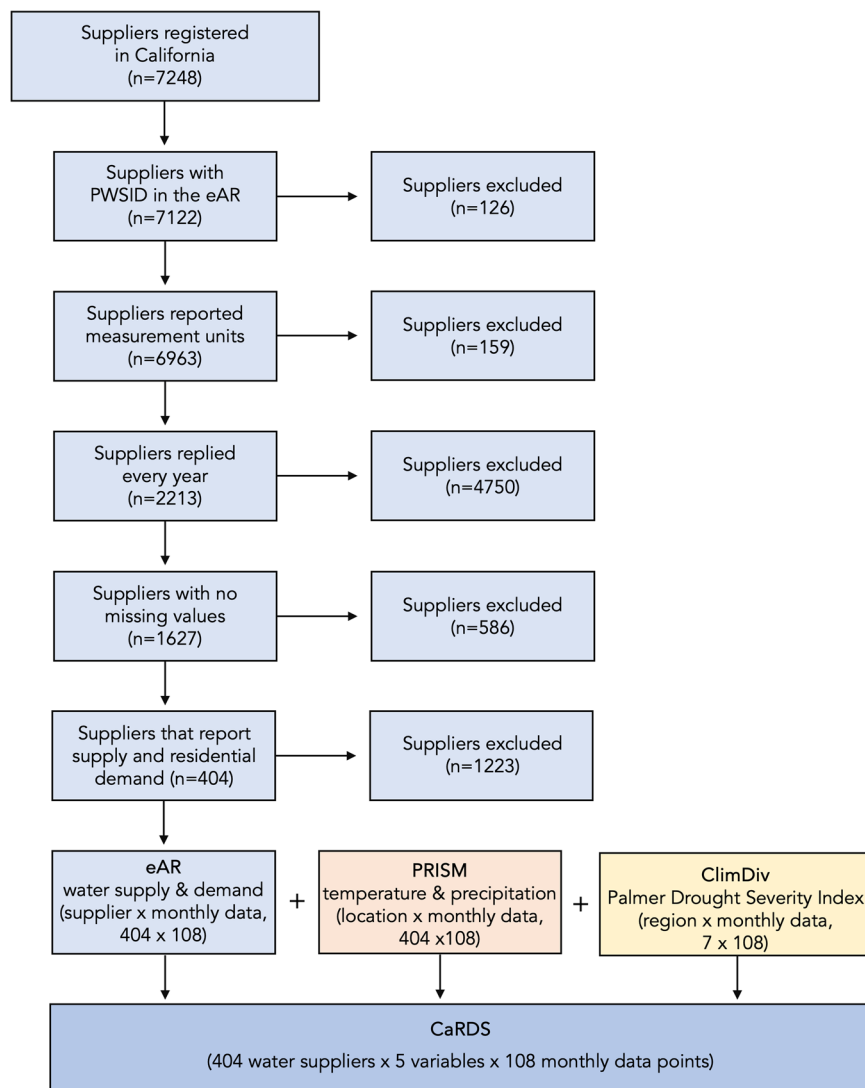


Fig. 1 Flow diagram with exclusion criteria and relative data records excluded/retained in the creation of CaRDS. The flow diagram reports the exclusion criteria applied to the data retrieved from the Electronic Annuals Report (eAR) from the State Water Resources Control Board. The flow diagram is adapted from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines (PRISMA Flow Diagram²⁵).

challenge is that the number of suppliers varies heavily between the different reporting years, ranging from 4000 to 7000 answers. To account for continuous reporting and create a reliable dataset with consistent multi-annual time series of data for each supplier, we apply the exclusion criteria shown in Fig. 1, where criteria 1 to 4 are evaluated separately on the water supply and demand time series. A supplier is excluded while compiling CaRDS if:

1. they do not report a unique identifier, i.e., the Public Water System Identification Number (PWSID), in their annual reporting.
2. they never report the unit for water supply and demand in the study period.
3. they do not report every year during the study period.
4. their report has one or more missing water supply or demand (total and residential) values during the study period.
5. they do not report water supply and water demand for residential use.

Missing or inconsistent information in other selected variables (i.e., measurement units for supply, measurement units for demand, and people served) is not a reason to exclude suppliers. Missing population data is interpolated by using the previously reported value, as population numbers reported by the water suppliers are very stable. The last step to create a consistent dataset is converting water supply and demand values into standard measurement units according to the International System of Units. Water suppliers often report different units for water supply and demand, thus we convert them separately. Additionally, starting in 2020 there is no information on measurement units for all water suppliers. We solve this problem by applying the following two unit conversion steps:

1. Compute the difference in magnitude to January 2019

$$\Delta_{i,N} = \frac{X_{i,Jan19}}{X_{i,JanN}}, \quad (1)$$

with X being the reported value for $i = \{\text{Supply, Demand}\}$ and $N = \{2020, 2021\}$.

2. Assign measurement units and convert to standard units.

$$\hat{X}_{i,n} = \begin{cases} X_{i,n}, & \text{if } \Delta_{i,N} < 600 \text{ (Values reported in Gallons),} \\ X_{i,n} * 748, & \text{if } \Delta_{i,N} 600 < 3,000 \text{ (Values reported in CCF),} \\ X_{i,n} * 325,851, & \text{if } \Delta_{i,N} 3,000 < 500,000 \text{ (Values reported in AF),} \\ X_{i,n} * 10^6, & \text{if } \Delta_{i,N} > 500,000 \text{ (Values reported in Mio. Gallons).} \end{cases} \quad (2)$$

with n being the monthly value we need to convert.

To ensure the unit conversion is successful, we analyze the temporal consistency of the time series by means of outlier detection (see Section "Consistency of time series and potential outliers").

PRISM Climate Data. To account for hydroclimatic influences and conditions on water management strategies, we include the monthly mean temperature and cumulative precipitation for the service area of each supplier in our dataset. The PRISM Climate Group¹⁸ gathers climatic data from various sources, applies quality control mechanisms, and releases various climate datasets with multiple spatial and temporal resolutions for the USA. The coordinates for each supplier's location are needed to obtain the related climatic data. Based on the ZIP code of a supplier we compute the spatial centroid for its service area and use the centroid coordinates as input for the PRISM data retrieval. This way we could match all supplier locations with the exception of two that were added manually. With the batch retrieval we compute a mean monthly temperature (in Celsius) and the cumulative monthly precipitation (in Millimeters) for each supplier location.

ClimDiv. Given the historical importance and environmental and socio-economic impacts of multi-year droughts in California, we include the Palmer Drought Severity Index (PDSI)²⁰ as an additional hydroclimatic factor in CaRDS. PDSI is a measure to estimate relative dryness and it is very effective in accounting for long-term drought conditions, taking the basic effects of global warming into account. The PDSI is provided by NOAA and calculated for large areas¹⁹, roughly following the division of hydrologic regions. NOAA divides California into seven climate divisions, instead of the ten regions considered by the SWB. This is achieved by either merging two hydrologic regions together (San Joaquin River and Tulare Lake; South Lahontan and Colorado River) or splitting one region before merging (Central Coast is split between San Francisco Bay and South Coast). We retrieve monthly PDSI values for each of the seven divisions for the study period of nine years and match them to each water supplier based on ZIP Codes.

Data integration and compilation of CaRDS. After data processing on the three individual datasets presented in the previous sections, we integrate them and compile CaRDS (see Fig. 1). Each time series we extract is linked to the unique PWSID, making it easy to merge the data and have a consistent set of water supply/demand and hydroclimatic variables for each supplier. The version of CaRDS released with this publication includes 404 water suppliers, each with five corresponding time series of the following monthly variables: water supply, water demand, mean temperature, cumulative precipitation, and PDSI. Each variable has a length of 108 time steps, covering in total nine years in monthly intervals. A detailed overview of the variables included in CaRDS, along with a short description and their units, is provided in Table 1.

Data Records

The CaRDS²¹ dataset is available on HydroShare and can be accessed via the following link: <https://doi.org/10.4211/hs.4ec7019fe63944bf87d40d2cdfa0d686>. The data is structured by two levels of key identifiers. The first level is the unique supplier identification number PWSID and the second level contains the time-series of monthly water supply, demand, mean temperature, cumulative precipitation, and PDSI (see Table 1). In the same repository we also share a file called *Supplier_Info.csv*, which provides secondary information about the suppliers in our dataset. This file mostly contains geographic information (ZIP code, county, hydrologic zone, climatic zone, and climatic division), along with information on the population served and the size of each supplier.

Technical Validation

As the CaRDS dataset we present here is largely based on survey data, using traditional approaches for data validation by modeling the retrieved data or comparing it to similar datasets is not possible. A way to check the validity and plausibility of the water supply and residential demand time series is to look at their patterns. In Fig. 2, we display the monthly distribution of the water supply and demand for all suppliers over the nine years included in CaRDS, as well as the computed daily per capita water use. For all three instances a distinct seasonal pattern emerges, with higher values during the summer periods. This behavior is expected as California overall

Variable name	Description	Unit
PWSID	Public Water System Identification Number	—
Supply	Monthly water produced by supplier	gal
Demand	Monthly water sold by supplier to residential customers	gal
Mean Temperature	Average monthly temperature in supplier location	°C
Precipitation	Cumulative monthly precipitation in supplier location	mm
PDSI	Palmer Drought Severity Index in supplier climatic division	—

Table 1. Overview of the variables included in CaRDS.

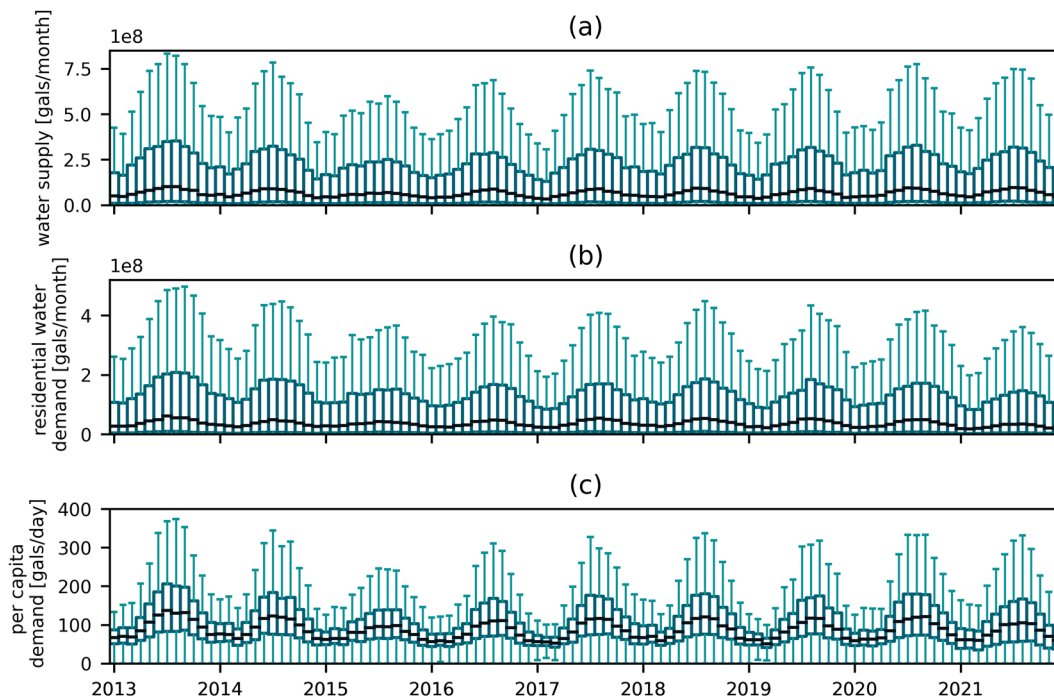


Fig. 2 Time series of (a) water supply, (b) residential water demand, and (c) daily per capita water consumption for all water suppliers included in CaRDS. The line in the middle of each box represents the median value across all suppliers, boxes mark the first and third quartile. The length of box whiskers is equal to 1.5 times the interquartile range. In all three instances a distinct seasonal pattern emerges, with higher values during the summers. This behavior is expected as California overall has wet winters and dry summers. Further, there is a noticeable smaller peak in all three instances in the summer of 2015. Mandatory water conservation measures were in place in California during this period to overcome the ongoing drought. Data in CaRDS thus capture climate-driven water management and policy dynamics.

has wet winters and dry summers. Further, there is a noticeable smaller peak in all three instances in the summer of 2015. Water scarcity and policy decisions resulted in establishment of mandatory water conservation measures in California to overcome the ongoing drought during that period. This implies that our dataset is able to reasonably and plausibly capture both the seasonal nature of its variables, and the influence of water management dynamics.

Consistency of time series and potential outliers. To consolidate and assess our data processing methodology, in particular the conversion of measurement units, we apply Tukey's fences²² separately for the water supply and demand time series of each supplier. By using interquartile ranges, we can identify possible and probable outliers for each time series. We find that 32% of the supply time series and 49% of the demand time series have possible outliers, while probable outliers exist in 9% and 17% of the supply and demand time series, respectively. These values are non-negligible. However, the distributions in Fig. 3 show that most water suppliers exhibit no outliers and an additional 4% (supply) and 20% (demand) exhibit only between 1 and 5 outliers in their monthly supply or demand values. We detect a small incline around 12 detected outliers in the cumulative distributions in Fig. 3 across all classified outliers for supply and demand. This can indicate that the measurement units might have been wrongly reported for one full year by only 10% of the suppliers. Possible and probable outliers in water supply and demand data are thus expected to only marginally influence the quality of data for individual

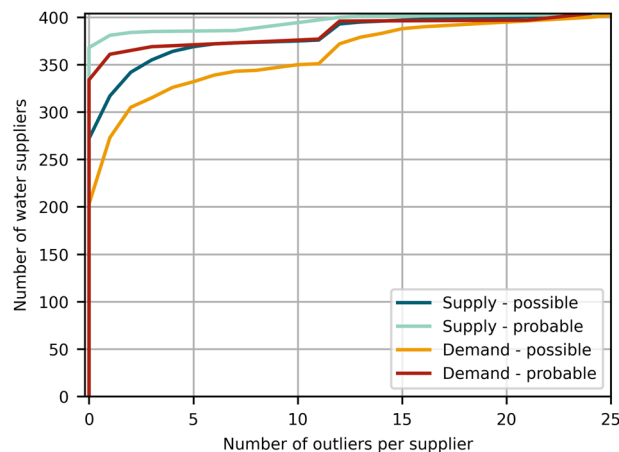


Fig. 3 Cumulative distribution of possible and probable outliers per water supplier by applying Tukey's fences to each water supply and demand time series in CaRDS.

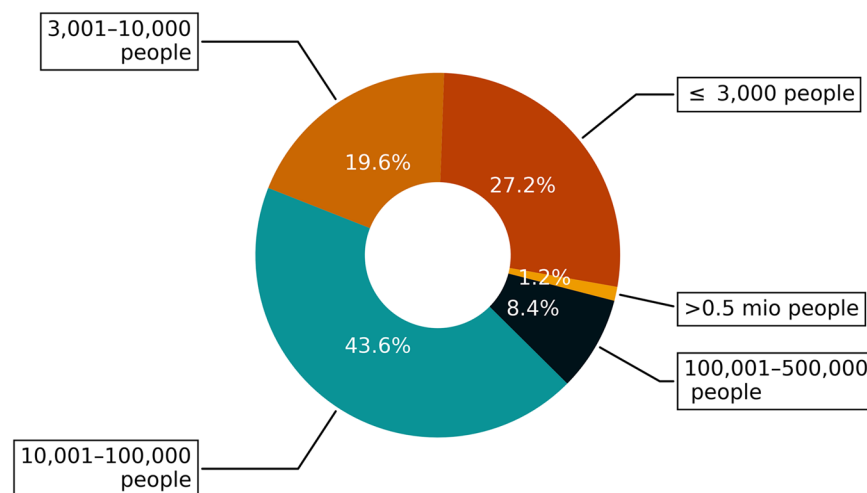


Fig. 4 Relative proportion of water suppliers included in CaRDS across different utility sizes. Utility size is quantified as the population served by each supplier.

suppliers (i.e., there are no suppliers with major portions of outliers in their data time series). Overall, the dataset encompasses two climatic extreme events, where outlier values are expected to a certain degree. Further, the computation of missing measuring units is not exact, but rather an approximation based on empirical value ranges. Nevertheless, a deviation in value magnitude between 2019 and 2020 is only detected for 4% of the suppliers underlining the validity of the approach.

Aiming for an automated and general way of pre-processing the different data sources without resource-intensive and subjective manual cleaning, and considering the qualitative challenges of the data collection, we decided not to exclude the potential outliers from the CaRDS dataset we are releasing with this publication. We thus did not exclude any further outliers from CaRDS after application of the initial exclusion criteria (Fig. 1), to preserve the original data structure and to give future users the possibility to rely on as many data as possible and optionally remove further data depending on their specific research needs. To further study the nature of each outlier and possibly remove some of them in case some applications based on CaRDS require it, an in-depth analysis of each time series may be necessary. More advanced outlier detection for time series data can rely on Autoregressive Integrated Moving Average (ARIMA)²³ models or unsupervised clustering such as DBSCAN²⁴, but further outlier detection is out of the scope of this study.

Analysis of population served. We analyze the population that is served by the water suppliers included in CaRDS to demonstrate that the CaRDS dataset is overall representative of the state of California. The water suppliers in CaRDS serve 52% (20. mio.) of the population in California. We further investigate the size of the communities that are served by the water suppliers (see Fig. 4). We see that nearly 50% of suppliers serve communities smaller than 10,000 people, representing small towns or neighborhoods and very rural settlements. The other

Climate Zone			Hydrologic Region		
Zone ID	Supplier count	Served population	Hydrologic region	Supplier count	Served population
1	15	174,159	North Coast	33	441,043
2	26	405,755	Sacramento River	35	954,708
3	29	5,654,892	North Lahontan	8	50,862
4	11	525,014	San Francisco Bay	26	6,221,284
5	22	333,704	San Joaquin River	36	912,626
6	20	901,753	Central Coast	50	1,026,902
7	5	451,599	Tulare Lake	24	1,041,544
8	33	2,774,847	South Lahontan	36	653,322
9	53	1,989,615	South Coast	140	8,603,622
10	31	2,668,344	Colorado River	16	214,241
11	16	348,472			
12	51	1,841,846			
13	21	1,084,999			
14	37	671,396			
15	9	172,470			
16	25	121,289			

Table 2. Distribution of water suppliers and corresponding population served in the 16 climate zones and 10 hydrologic regions in California. The climatic zones are based on the zones defined by the Building Energy Efficiency Standards by the California Energy Commission.

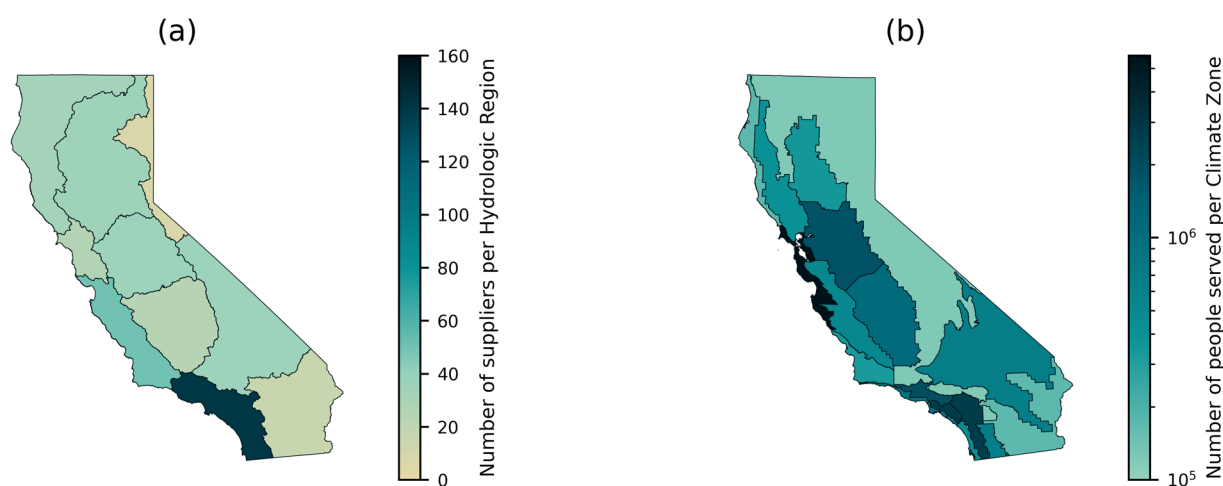


Fig. 5 Spatial distribution of water suppliers included in CaRDS. The count of water suppliers located in each of the 10 Hydrologic Regions in California is reported in (a). The amount of people served by these suppliers in each of the 16 Climate Zones is represented in (b).

half of the suppliers serve medium-sized towns and big cities, representing the more urbanized and metropolitan areas of California. CaRDS therefore represents well small and medium-sized suppliers, while large suppliers, e.g., Urban Water Retail Suppliers (URWS), are underrepresented. If the aim is to investigate URWS only, other data sources such as the SWB water conservation portal may be more suitable. A detailed overview of how many people are served by hydrologic region and climatic zone in California can be found in Table 2.

Spatial analysis. To further verify the spatial representation of CaRDS, we present different spatial distributions of the suppliers in the state of California. To demonstrate that the dataset achieves a satisfactory representation of water suppliers in California, Fig. 5(a) shows the number of suppliers of CaRDS in each of the 10 hydrologic regions California is divided in. We see that, first, each hydrologic region is represented in the dataset. Second, urbanized areas of the state are reflected with a higher number of suppliers being in metropolitan regions (South Coast and San Francisco Bay), and fewer suppliers in rural areas or areas further from metropolitan centers and core infrastructure (North Lahontan and Colorado River). Figure 5(b) shows a similar spatial distribution for the amount of customers served by the suppliers in CaRDS based on the 16 climatic zones in California. A detailed overview of the number of suppliers and population served per hydrologic region and climate zone can be found in Table 2.

Usage Notes

Data users should take into account the assumptions we made in creating the dataset. For water supply, eAR records do not directly report the monthly water that is supplied to customers, but the water that is produced in a month. Water production means water is either treated surface or ground water, bought from another supplier, or reused and then introduced into the water suppliers' network through direct distribution or storage. The storage component means that the data might include time lags for when water is actually supplied. Further, the eAR does not report the actual demands of households, but the amount of water that suppliers billed to residential customers. This means that the recorded water demand is dependent on the water meter/reading resolution and meter reading accuracy. In California, most larger urban water suppliers have metered data, but some smaller suppliers may use non-metering methods to quantify customer water use. As socio-economic factors become more important for water management research, the provided supplementary information of the suppliers (see Data Records) gives information on the population served by each supplier. Further, the provided ZIP codes can be used to cross-correlate the data in CaRDS with those from the U.S. Census Bureau. There is a wealth of socio-economic data already available at different spatio-temporal resolutions and it is well organized in their public available data repository ([census.gov](https://www.census.gov)).

Code availability

The data pre-processing leading to the development of CaRDS is based on open source Python software. Jupyter Notebooks with the code to pre-process, transform, and merge the different data sources reported in this article are available on HydroShare²¹ at <https://doi.org/10.4211/hs.4ec7019fe63944bf87d40d2cdfa0d686>.

Received: 15 December 2023; Accepted: 4 June 2024;

Published online: 14 June 2024

References

- Zhang, F. *et al.* Five decades of observed daily precipitation reveal longer and more variable drought events across much of the western United States. *Geophysical Research Letters* **48**, e2020GL092293 (2021).
- Greve, P. *et al.* Global assessment of water challenges under uncertainty in water scarcity projections. *Nature Sustainability* **1**, 486–494 (2018).
- Vicuna, S., Maurer, E. P., Joyce, B., Dracup, J. A. & Purkey, D. The sensitivity of California water resources to climate change scenarios 1. *JAWRA Journal of the American Water Resources Association* **43**, 482–498 (2007).
- Fu, X. & Tang, Z. Planning for drought-resilient communities: An evaluation of local comprehensive plans in the fastest growing counties in the US. *Cities* **32**, 60–69 (2013).
- Furlong, C., Brothie, R., Considine, R., Finlayson, G. & Guthrie, L. Key concepts for integrated urban water management infrastructure planning: lessons from Melbourne. *Utilities Policy* **45**, 84–96 (2017).
- Mitchell, D. *et al.* Building drought resilience in California's cities and suburbs. *Public Policy Institute of California* 1–49 (2017).
- Cahill, R. & Lund, J. Residential water conservation in Australia and California. *Journal of Water Resources Planning and Management* **139**, 117–121 (2013).
- Hanak, E. *Managing California's water: From conflict to reconciliation* (Public Policy Instit. of CA, 2011).
- Quinn, T. Forty years of California water policy: What worked, what didn't and lessons for the future (2019).
- California State Assembly. Assembly bill no. 1668. https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB1668 (2018).
- California State Senate. Senate bill no. 606. https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB606 (2018).
- California State Water Resources Control Board. Water conservation portal. https://www.waterboards.ca.gov/water_issues/programs/conservation_portal/conservation_reporting.html Accessed on 21.11.2023 (2023).
- Wilkinson, M. D. *et al.* The fair guiding principles for scientific data management and stewardship. *Scientific data* **3**, 1–9 (2016).
- Zounemat-Kermani, M. *et al.* Neurocomputing in surface water hydrology and hydraulics: A review of two decades retrospective, current status and future prospects. *Journal of Hydrology* **588**, 125085 (2020).
- Daniel, I. *et al.* A survey of water utilities' digital transformation: drivers, impacts, and enabling technologies. *npj Clean Water* **6**, 51 (2023).
- Stagge, J. H. *et al.* Assessing data availability and research reproducibility in hydrology and water resources. *Scientific data* **6**, 1–12 (2019).
- California State Water Resources Control Board. Electronic annual report. https://www.waterboards.ca.gov/drinking_water/certificatedrinkingwater/ear.html Accessed on 21.04.2023 (2023).
- PRISM Climate Group, Oregon State University. PRISM time series data. <https://prism.oregonstate.edu> Accessed on 05.07.2023 (2023).
- National Oceanic and Atmospheric Administration. Historical palmer drought severity indices. <https://www.ncei.noaa.gov/pub/data/cirs/climdiv/> Accessed on 27.06.2023 (2023).
- Palmer, W. C. *Meteorological drought*, vol. Res. Paper No.45 (US Department of Commerce, Weather Bureau, 1965).
- Gross, M., Escrivá-Bou, A., Porse, E., & Cominola, A. CaRDS - the statewide California residential water demand and supply open dataset, *HydroShare*, <https://doi.org/10.4211/hs.4ec7019fe63944bf87d40d2cdfa0d686> (2024).
- Tukey, J. W. *Exploratory data analysis* (Addison-Wesley Publishing Company, 1977).
- Box, G. E., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. *Time series analysis: forecasting and control* (John Wiley & Sons, 2015).
- Ester, M. *et al.* A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, vol. 96, 226–231 (1996).
- Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic reviews* **10**, 1–11 (2021).

Author contributions

M.-P.G., A.E.-B., E.P., A.C. conceived the research. M.-P.G. conducted data acquisition, integration, and compilation. M.-P. G. wrote the software code for data pre-processing, analysis, and designed the visual elements of the paper. M.-P. G. and A.C. analyzed the results. M.-P. G. drafted the initial manuscript. All authors edited and reviewed the manuscript, and all approved the final version of the paper.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.-P.G.

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

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



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

© The Author(s) 2024