The Canadian Surface Prediction Archive (CaSPAr)

A Platform to Enhance Environmental Modeling in Canada and Globally

Juliane Mai, Kurt C. Kornelsen, Bryan A. Tolson, Vincent Fortin,
Nicolas Gasset, Djamel Bouhemhem, David Schäfer, Michael Leahy,
François Anctil, and Paulin Coulibaly

ABSTRACT: The Canadian Surface Prediction Archive (CaSPAr) is an archive of numerical weather predictions issued by Environment and Climate Change Canada. Among the products archived on a daily basis are five operational numerical weather forecasts, three operational analyses, and one reanalysis product. The products have hourly to daily temporal resolution and 2.5–50-km spatial resolution. To date the archive contains 394 TB of data while 368 GB of new data are added every night. The data are archived in CF-1.6-compliant netCDF-4 format. The archive is available online (https://caspar-data.ca) since June 2017 and allows users to precisely request data according to their needs, that is, spatial cropping based on a standard shape or uploaded shapefile of the domain of interest and selection of forecast horizons, variables, and issue dates. The degree of customization in CaSPAr is a unique feature relative to other publicly accessible numerical weather prediction archives and it minimizes user download requirements and local processing time. We benchmark the processing time and required storage of such requests based on 216 test scenarios. We also demonstrate how CaSPAr data can be employed to analyze extreme rainfall events. CaSPAr provides access to data that are fundamental for evaluating numerical weather prediction models and demonstrating the improvement in products such as flood and energy demand forecasting systems.

AFFILIATIONS: Mai and Tolson—University of Waterloo, Waterloo, Ontario, Canada; **Kornelsen* and Coulibaly**—McMaster University, Hamilton, Ontario, Canada; **Fortin, Gasset, and Bouhemhem**—Environment and Climate Change Canada, Dorval, Quebec, Canada; **Schäfer**—Helmholtz Centre for Environmental Research, Leipzig, Saxony, Germany; **Leahy**—Esri Canada, North York, Ontario, Canada; **Anctil***—Université Laval, Quebec City, Quebec, Canada
* CURRENT AFFILIATION: Kornelsen—Ontario Power Generation, Niagara-on-the-Lake, Ontario, Canada

E nvironmental models such as hydrologic and land surface models require meteorological input data like precipitation and temperature to estimate, for example, discharge and soil moisture. In a hindcast period these data are usually provided by ground observations. When these models are, however, run in a forecast mode, they require forecasted meteorological inputs. These so-called numerical weather predictions (NWPs) are derived by atmospheric models.

The forecasts of, for example, river discharge produced by the hydrologic and land surface models using NWPs are important for, for example, flood and drought predictions (Xuan et al. 2009; Yucel et al. 2015; Cao and Zhang 2016; Rogelis and Werner 2018), optimal strategies for reservoir management (Yang and Yang 2014; Schwanenberg et al. 2015; Ficchì et al. 2016), and power production (Taylor and Buizza 2003; Soman et al. 2010; Foley et al. 2012; Ren et al. 2015).

To evaluate or improve such forecasting systems, archived NWPs are required to properly test new systems and demonstrate the improvements compared to previous system setups (Abaza et al. 2013; Thiboult et al. 2017). Data are usually disseminated by the agency producing the NWPs such as the European Centre for Medium-Range Weather Forecasts (ECMWF), the National Oceanic and Atmospheric Administration (NOAA) of the U.S. Department of Commerce, and Environment and Climate Change Canada (ECCC).

Archives of NWPs are, however, rare and for those that do exist, users often need to download the full dataset since a spatial, temporal, and/or variable selection is not possible. A notable exception of a data archive is NOAA's National Operational Model Archive and Distribution System (NOMADS) (Rutledge et al. 2006). This system is a repository of weather model output datasets, model input datasets, and a limited subset of climate model datasets generated by NOAA (www.ncdc.noaa.gov/data-access/model-data). In NOMADS the user can select variables, time periods, and even a bounding box (for selected products). The latter is, however, only possible using the Thredds Data Server, which was discontinued in 2018. The other download options, FTP and HTTP(S), provide only very short archives of a few days and only allow users to download the full data product without specifying the actual domain or variable of interest. NOMADS also offers HTTP adaptive streaming (HAS) for several products, which gives access to longer archives but also only allows the download of the complete domain, variable, and forecast. The NOMADS web page mentions a Network Common Data Form (netCDF) subsetting service, which might be employed to customize a data request. Spatial domain selection, however, would be limited to rectangular bounding boxes. The netCDF subsetting service lacks a graphical user interface and so this limits the utility and access of this service to potential users.

Another archive of (mostly) European products is distributed by the ECMWF (https://confluence.ecmwf.int/display/WEBAPI/Access+ECMWF+Public+Datasets). The ECMWF disseminates not only their own data but also several additional datasets including the THORPEX Interactive Grand Global Ensemble (TIGGE) database and a product contributed by ECCC that is upscaled to 100 km. The user has to specify manually the variable(s), time period, ensemble size, issue date(s), and vertical level for the desired data product while a spatial selection of the domain of interest is not possible. The data are provided then as a download link when the processing has finished. The ECMWF is also hosting the Climate Data Store (https://climate.copernicus.eu/climate-data-store), which is an online tool that allows users to browse available products and supports building applications using these data.

An archive of NWPs issued by ECCC is not yet publicly accessible. ECCC provides operational NWPs via the Datamart system (https://dd.weather.gc.ca/about_dd_apropos.txt). Datamart holds data of the last few days for only a limited set of variables and hence is not serving as an archive. The data are provided in files for each forecast horizon, variable, issue date, and ensemble member without allowing for spatial subsetting of the data. The data are provided in grib2 format, which required a spatial interpolation of the original NWPs onto a grid supported by grib2.

The absence of an archive of NWP data produced by ECCC, as well as the limited existing archives that have limited or no spatial and temporal subsetting, led to the development of the Canadian Surface Prediction Archive (CaSPAr). CaSPAr is accessible free of charge and archives eight operational NWPs, one preoperational product, and one reanalysis dataset that are all generated by ECCC at hourly to daily temporal resolution and 2.5–50-km spatial resolution. CaSPAr provides a web-based GIS platform that allows the user to select variables, time period and forecast horizons of interest as well as the spatial subsetting of the products beyond bounding boxes: the domain of interest can either be selected by drawing a polygon (by hand) or uploading a shapefile. A description can be found in the wiki associated with the CaSPAr GitHub (https://github.com/julemai/CaSPAr/wiki/How-to-get-started-and-download-your-first-data). The data provided in CaSPAr are not interpolated (as in Datamart or ECCC product distributed by ECMWF). CaSPAr provides access to forecasts that were issued seven days ago or earlier meaning that it does not provide the current forecasts. This is due to the fact that the CaSPAr system is decoupled from the data provider (ECCC) and ensures that all data were retrieved and checked for consistency. It also ensures there is no confusion with the source for operational data. The data requested by a user are processed on the back end and the user is notified by email with a link to download the data. The download is realized using Globus (Foster 2011; Allen et al. 2012), which allows for secure and fast, parallel data transfers that is free of charge for the users.

A core philosophy of CaSPAr is user data sovereignty. This means that users should ultimately determine how their data should be represented for their unique modeling and analysis needs. Unlike other data providers, CaSPAr stores all data on the original grid used by the NWP model without any interpolation or reprojection. This philosophy provides transparency for CaSPAr users and provides them the power to choose which transformations and algorithms are applied to their data with full knowledge of the strengths and limitations therein.

In the following, we will introduce the structure of the CaSPAr framework and ECCC's NWP products archived in CaSPAr (both in the second section). A collection of test requests is performed to demonstrate the processing times and request storage sizes that will allow the readers to extrapolate the waiting times and storage requirements for their own requests (third section). A record flood event for southern Ontario, Canada, that occurred in June 2017 will be revisited to visualize the various precipitation forecasts that can be drawn from CaSPAr. The example is also used to determine how much the size of a download can be reduced by

specifying exactly the required data (fourth section). In the fifth section, we will provide concluding remarks and next steps to further improve CaSPAr.

Material and methods

We will first give an overview of the general structure of the CaSPAr to provide insights into its functioning and capabilities ("Components of the CaSPAr framework" section). This is followed by a brief introduction of the NWP and reanalysis products that are currently archived in CaSPAr ("Data products" section). All archived data follow the same file naming convention and netCDF structure ("Data structure" section).

Components of the CaSPAr framework. The CaSPAr framework can generally be divided into two components—the front end and the back end. Both are based primarily on Python and bash scripts. The front end delivers the web interface (at www.caspar-data.ca), which supports the platform for user registration (→Register), contact information (→Contact), a link to CaSPAr's documentation and pre/postprocessing scripts on GitHub (→Documentation), a catalog to get an overview of products, variables, forecast lengths and periods where data are available (→Catalogue), and most important the interface that allows the users to request data (→Data Portal). The front end holds a metadatabase of all data available in the archive but not the data itself. This means the front end knows that a forecast exists for a specific issue day of an individual variable but it does not know the forecast itself. The front end also holds all the user relevant information required to allow users to access CaSPAr. Using Esri's ArcGIS Enterprise (https://enterprise.arcgis.com/), the graphical interface for the front end was built with Portal for ArcGIS. Custom geoprocessing tools that support the interface and provide an application programming interface (API) for advanced CaSPAr users were built with ArcGIS Server. User requests submitted to the front end are queued and sent to the back end via a secure connection.

The back end does not store user or request history and instead only processes requests the front end has sent and informs the front end once processing has finished. The front end then sets the file permissions to allow the user to access the individually processed data and notifies the users that the processing has finished. The notification is done by email containing a download link. The download is conducted using the grid FTP method managed by Globus. Globus is free of charge for the end users and provides secure, resumable, and parallel file downloads. Globus can also be accessed through an API allowing users to fully automate their interactions with CaSPAr. The Globus service is provided by Compute/Calcul Canada.

Every night the back end server pulls new NWP data directly from ECCC's servers. The data received are in ECCC's internal, binary file format called FST that can only be read using a ECCC internal library limiting the readability and usability of the data. They are hence converted into a more widely known format: a CF-1.6-compliant netCDF-4 format. The netCDF format was chosen as it is platform independent and self-describing, meaning the file contains all the necessary metadata information to understand the content of the file. Following the Climate and Forecast (CF) convention is beneficial since it eases the ingestion of these files in many processing libraries. The conversion from FST into netCDF includes the proper setting of metadata for all variables, encoding of projection information from the FST format and setting projection variables in the netCDF file. Missing or inconsistent data are reported to the database administrator. Once new netCDF data are successfully pulled into the CaSPAr database, the back end registers the metadata in the front end's metadatabase. Every night about 368 GB of new data are processed and added to the database. To date, the database contains 395 TB of NWPs in netCDF format (as of February 2020).

Data products. Currently, CaSPAr archives eight operational products, that is, the Global Ensemble Prediction System (GEPS) (Charron et al. 2009; Houtekamer et al. 2014; Gagnon et al. 2015; Lin et al. 2016; CMC 2018f), the Global Deterministic Prediction System (GDPS) (CMC 2018e), the Regional Ensemble Prediction System (REPS) (Lavaysse et al. 2013; CMC 2018d), the Regional Deterministic Prediction System (RDPS) (CMC 2018g), the High-Resolution Deterministic Prediction System (HRDPS) (Milbrandt et al. 2016; CMC 2018b), the Canadian Land Data Assimilation System (CaLDAS) (Balsamo et al. 2007; Carrera et al. 2009), and the Canadian Precipitation Analysis on a coarse 10-km grid (CaPA coarse) (Mahfouf et al. 2007; Carrera et al. 2009; Lespinas et al. 2015; Fortin et al. 2015, 2018; CMC 2018c) and on a finer 2.5-km grid (CaPA fine) (Fortin et al. 2018; CMC 2018a). CaSPAr also hosts the finer CaPA product in its preoperational version (CaPA fine exp.) and a reanalysis of RDPS called the Regional Deterministic Reanalysis System (RDRS). The archive starts for most products in May 2017. Exceptions are the RDRS product (available from January 2010 to December 2014; will be extended soon to cover January 1980 to December 2018), the RDPS product (available since January 2015), the CaPA at coarse resolution (available since September 2012) and the preoperational CaPA at fine resolution (available from June 2016 until it became operational in March 2018).

The spatial extent of the products can be found in Fig. 1. GEPS and GDPS cover the whole globe while REPS, RDPS, RDRS, and CaPA coarse are available for North America and HRDPS, CaLDAS, and CaPA fine (all 2.5-km resolution) cover Canada and northern parts of the United States. For details on spatiotemporal resolution, number of issues per day, number of variables, forecast horizons, and ensemble size (users should refer to Table A1 in the appendix). The information are subject of change since ECCC constantly updates and improves the operational NWPs. To get the most up-to-date information the users should visit our GitHub documentation (https://github.com/julemai/CaSPAr/wiki/Available-products) as well as the CaSPAr Data Catalogue (www.caspar-data.ca, →Catalogue). In the Data Catalogue, the user can select a product and will receive a string that contains all the product specific details like available variables and forecast horizons. The information returned by the Data Catalogue will be required when using the API rather than the web-based front end. For details on the products themselves and their production, users should refer to ECCC's documentation. A list of product-dependent documentations is available (under https://dd.meteo.gc.ca/about_dd_apropos.txt).

Please note that the data are provided with a 7-day delay. This time is used to ensure that the data retrieval from ECCC, the conversion of data into netCDF and the updating of the CaSPAr database are successful.

Data structure. The data themselves fulfill the CF-1.6 netCDF convention. All files follow a standard naming pattern, which is YYYYMMDDHH.nc for deterministic and YYYYMMDDHH_EEE.nc for ensemble products, where YYYY is the year, MM is the month, DD is the day, and HH is the hour of the product issue date and time in UTC. EEE indicates the ensemble member. The numbering of ensemble members starts with zero as the control member.

The naming pattern already indicates that each file always contains the whole forecast including all forecast horizons and all variables at all horizontal levels. Ensemble members are in separate files. The separation of ensemble members in different files, allows every forecasted variable to be stored on three-dimensional arrays (latitude, longitude, time). This eases the usage of those files as direct inputs for models and the postprocessing of files in general. As an example, one would retrieve 126 files when requesting both issues of the REPS product with 21 ensemble members over a 3-day period.

All products are provided by ECCC on native rotated latitude–longitude grids. Data were kept on NWP model grids to avoid spatial interpolation and maintain the philosophy of user data sovereignty. The (spatial) dimension names in the netCDF are rlat and rlon. The
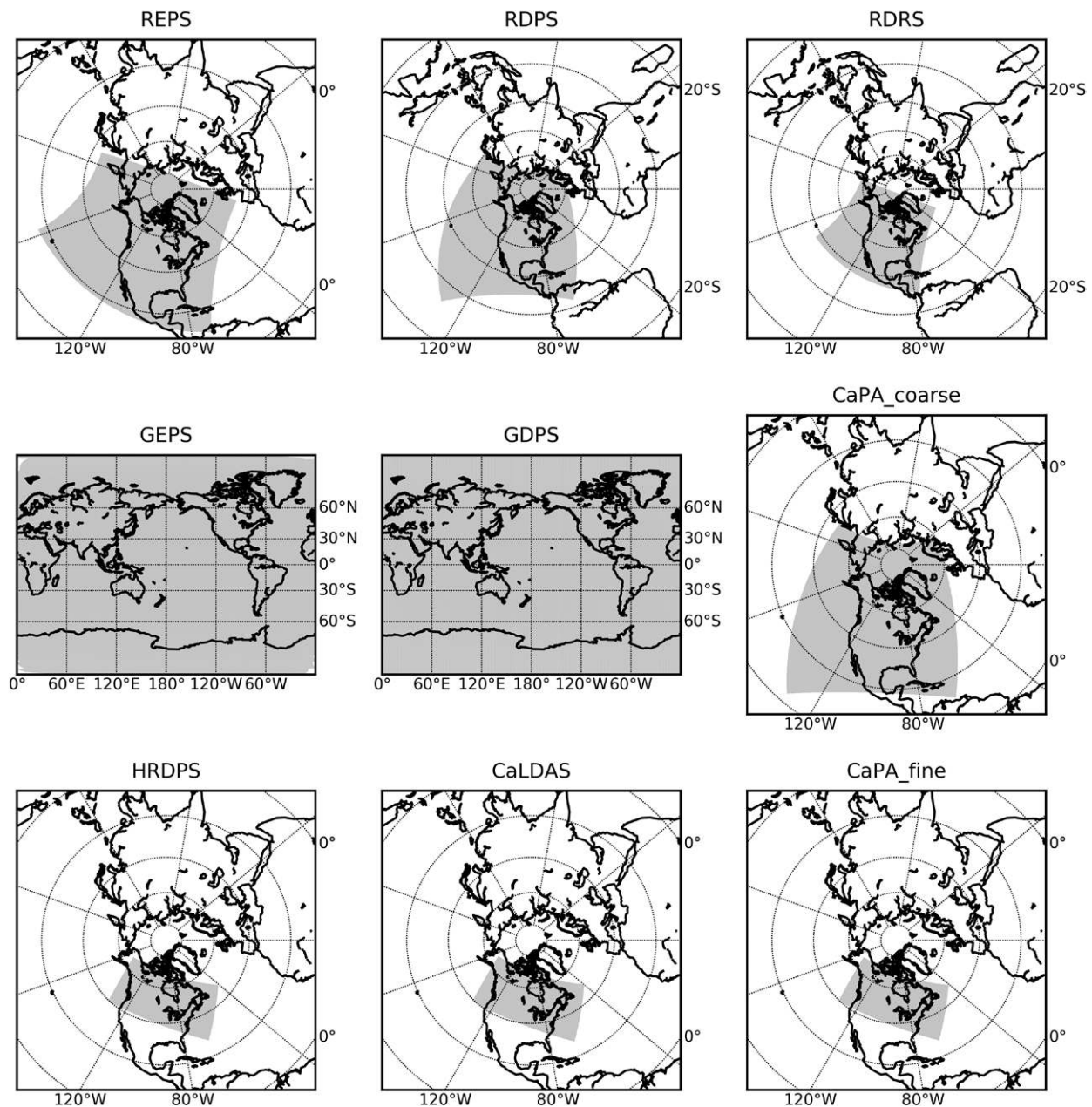
Fig. 1. The spatial extent of nine products available in CaSPAr is highlighted in gray. The domains are expected to change over time since the spatial resolution is continuously improved. GDPS (25 km) and GEPS (50 km) are global products. RDPS (10 km), REPS (15 km), RDRS (15 km), and the coarse (10 km)-resolved CaPA cover North America. The products with the highest spatial resolution of 2.5 km, that is, HRDPS, CaLDAS, and CaPA_fine, cover Canada and northern part of the United States.

one-dimensional coordinates (before transformation) are named rlat(rlat) and rlon(rlon). The variable rotated_pole contains the projection information. The transformed (rotated) two-dimensional latitude and longitude grids are added for convenience as lon(rlat,rlon) and lat(rlat,rlon). They are derived by applying the projection information on the one-dimensional coordinates. The only exception to the user data sovereignty philosophy is that the UU and VV components of wind are provided in the native version, where the wind components are along the rotated grid axis, and in a derived format, where the wind components are given along the geographical latitude and longitude axis.

Run time and storage benchmark of CaSPAr data

To benchmark the run times and storage requirements of potential user-specific data requests of different complexity, we processed 216 test requests for the nine available products, six domains of interest, and four different scenarios. The requests were submitted using the web front end and are documented and described in detail in the GitHub wiki (https://github.com /julemai/CaSPAr/wiki/How-to-get-started-and-download-your-first-data). As the benchmark results show below, processing times can range from minutes to multiple days and storage requirements can be up to 80 GB depending on the user request. The detailed benchmark results below provide users with a way to estimate and plan for potentially long wait times and very large storage requirements associated with their specific requests.

The six domains used for the benchmark were chosen such that their areas cover a range of several orders of magnitudes. All watersheds are located in Canada to allow for test requests of all products. The domains are the Grand River draining to West Montrose and Port Maitland (1,148 and 6,702 km^2, respectively), the Madawaska watershed (8,501 km^2), Red Deer watershed (11,611 km^2), Lake Erie basin including Lake Saint Clair (103,666 km^2), and the entire Hudson Bay (5,204,940 km^2). The latter has a land area of 3.86 million km^2, which corresponds to more than 42% of the Canadian land area. It can hence be regarded to be one of the largest domains a user would request. The results of the benchmark are independent of the location of the domains (readers interested in the benchmark domains can see them mapped in Fig. ES1 of the online supplemental material; https://doi.org/10.1175/BAMS-D-19-0143.2). The six domains are submitted as shapefiles to CaSPAr leading to the smallest possible file size since only data relevant for the domain of interest are stored while the outer areas are filled with no-data values. No-data values do not increase the size of netCDF-4 files.

The different spatial resolutions of the products and the varying sizes of the domains lead to different amounts of grid cells that cover the domain of interest. The number of grid cells extracted are listed in Table 1. The coarsest product, GEPS (50 km) leads to the smallest number of grid cells (e.g., 4,417 for Hudson Bay). HRDPS, CaLDAS, and CaPA (fine) all have approximately a 2.5-km resolution but are on slightly different rotated grids. This results in 793,342 (HRDPS and CaPA) and 799,810 (CaLDAS) extracted grid cells for the Hudson Bay.

The four benchmarked scenarios with increasing complexity have been tested. In the first scenario (scenario A) only one issue of one day and one variable is requested. Scenario B is the same as scenario A but now all forecasted horizons are requested. Scenario C is the same as scenario B but it requests 30 days of forecasts instead of only 1 day. Scenario D is the same request as scenario C but now all variables are requested instead of only one. The benchmark tracked both processing time, that is, the time a user would need to wait until he/she receives the email with a download link, and storage requirements of the processed request, that is, the storage the user needs to have available when downloading the data.

For the benchmark scenarios A–C only variables are selected that are available at all forecasted time steps (horizons) and over the whole time period to avoid an underestimation of storage measures. Real-world requests might be smaller in size due to the fact that no-data values do not increase the file size in netCDF-4 format.

Table 1. Area of the six domains tested for requesting the nine available products in CaSPAr. The different spatial resolutions of the nine products lead to specific amounts of grid cells that are extracted when data of the respective domain is requested. The Global Ensemble Prediction System (GEPS) is the coarsest product (approximately 50 km at the equator) while the High-Resolution Deterministic System (HRDPS), the Canadian Land Data Assimilation System (CaLDAS), and the Canadian Precipitation Analysis (CaPA fine) are the finest (approximately 2.5 km). CaLDAS is on a slightly different rotated grid compared to CaPA fine and HRDPS, which leads to minor differences in the number of grid cells.

Domain name	Area (km²)	No. of grid cells								
		REPS	RDPS	RDRS	GEPS	GDPS	CaPA coarse	HRDPS	CaLDAS	CaPA fine*
Grand River (WM)	1,148	47	58	26	7	8	67	322	322	322
Grand River (PM)	6,702	118	177	80	16	24	196	1,443	1,443	1,443
Madawaska	8,501	136	207	91	21	31	233	1,800	1,798	1,800
Red Deer	11,611	143	243	105	26	39	259	2,313	2,311	2,313
Lake Erie	103,666	797	1,539	666	109	213	1,605	18,529	18,527	18,529
Hudson Bay	5,204,940	24,469	54,821	23,640	4,417	10,714	55,271	793,342	799,810	793,342

*The operational CaPA product at the fine resolution (CaPA fine) is the same as the preoperational product (CaPA fine exp.).

The results of the 216 benchmark requests are shown in Fig. 2. Please note that the *y* axes in the figure are all in logarithmic scale. We will first analyze the storage requirements (colored bars) and then the processing times (gray bars). The precise values of the estimates for all 216 requests can be found in the supplemental material (in Table ES1).

REPS, GEPS, and HRDPS lead to most storage intense requests. The request scenario D over the Hudson Bay (yellow bars in Fig. 2, first column of panels) led to a storage size of 79.7, 60.5, and 66.3 GB, respectively. The storage size for RDPS, GDPS, and CaLDAS are in the range of a few gigabytes; for RDRS and CaPA fine it is slightly less than 1 GB. The storage requirement for CaPA coarse is a couple of megabytes. Requesting all variables (scenario D; yellow bars) versus one variable (scenario C; light green bars) is almost same for both CaPA products since they contains only two variables in total. Requesting all horizons (scenario B; dark green bars) versus one horizon (scenario A; dark blue bars) is exactly the same for the CaPA products since they are analysis products (not forecast) and hence contain only one time step. The storage requirements are almost same for the four smallest domains (all less than 12,000 km²).

The processing time reported for comparison is CPU clock time as reported by the Compute Canada Graham Cluster. The value reported is hence the time the user has to wait to get a download link assuming that there is no queue on the cluster. CaSPAr does not yet take advantage of the parallel computing capabilities of the cluster to process user requests but the processing times could be decreased by enabling this feature. The processing (gray bars in Fig. 2) surprisingly seems to be independent of the size of the domain of interest. Only the processing of requests for the Hudson Bay usually takes longer. The average processing time for the largest request scenario D (lightest gray bars in Fig. 2) of GEPS is around 37 h, REPS is around 11 h, GDPS is around 3.5 h, HRDPS is around 3 h, RDPS is around 2.5 h, CaLDAS is around 20 min, RDRS and CaPA fine are both around 1.5 min, and CaPA coarse takes around 30 s.

It should be noted that there is an upper limit of 1-TB storage and 28 days of processing time for requests. The user would receive an error message when he/she tries to submit a larger request.

Postevent analysis example using archived CaSPAr data

On 23 June 2017 an extreme rainfall was recorded in the Grand River watershed in Ontario, Canada. As reported by the local flood forecasting authority [Grand River Conservation
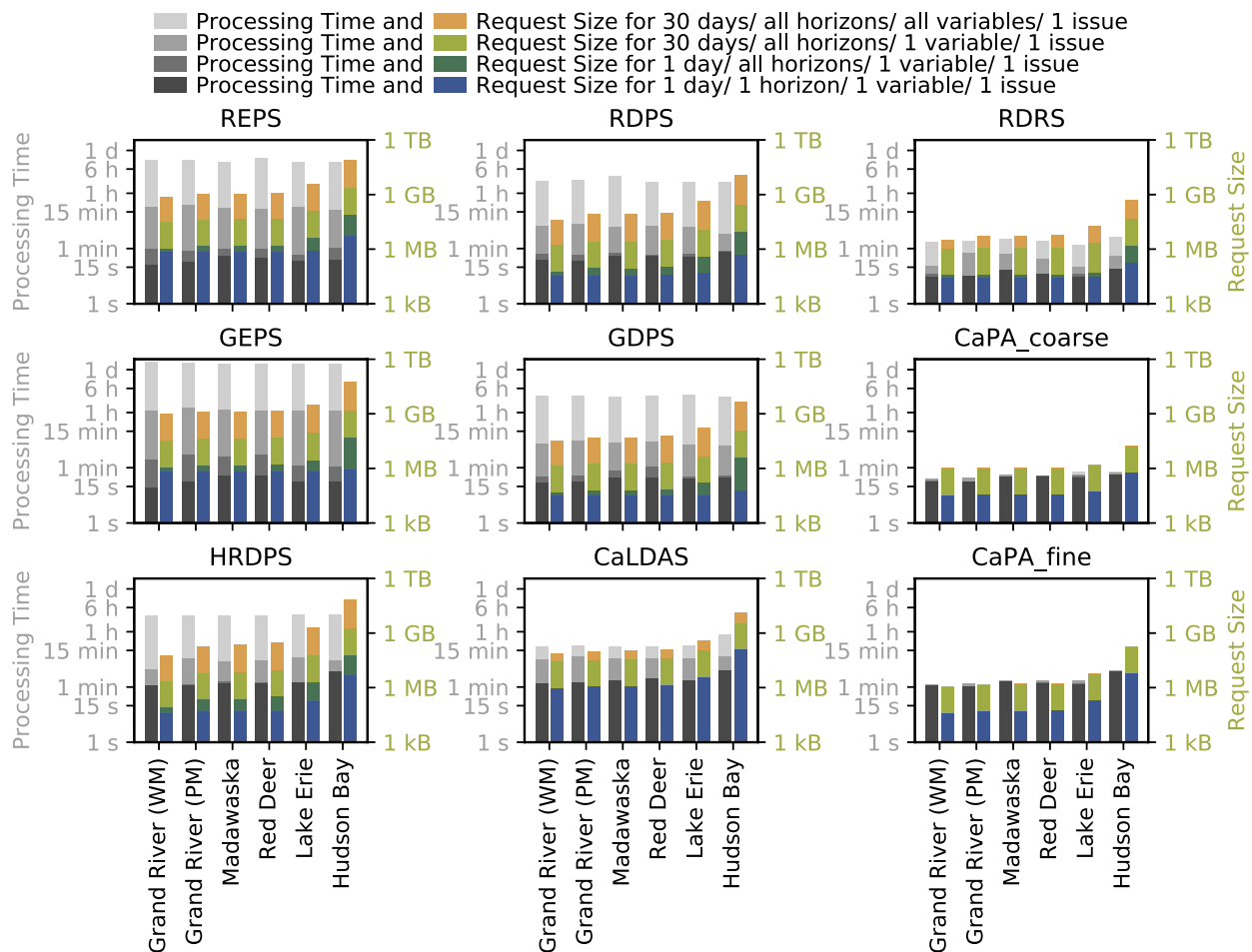
Fig. 2. CPU processing times (gray bars) and required storage (colored bars) for four different request scenarios from requesting only one issue of one day for one variable and one forecasted horizon (darkest gray bar for timing and blue for storage) to requesting one issue for 30 days of all forecasted time steps and all available variables (lightest gray for processing time and yellow for storage requirement). Since the four request scenarios are strictly increasing in complexity, the bars are plotted on top of each other. The data were requested for six different domains using their exact shape via shapefile upload in CaSPAr: Grand River watershed draining to gauge station West Montrose (WM) and Port Maitland (PM), Madawaska, Red Deer, Lake Erie (including Lake Saint Clair) watershed, and the entire Hudson Bay watershed, whose drainage area covers about 42% of all of Canada. Details on the size of the watersheds and the number of requested grid cells can be found in Table 1. The spatial location of the smallest domain [Grand River (WM)] and the according requested grid cells can be also found in the maps shown in Fig. 3. Please note that both *y* axes are in logarithmic scale. The exact values of the estimates can be found in the online supplemental material (in Table ES1).

Authority (GRCA); see www.grandriver.ca/en/our-watershed/Record-Rainfall-Flood-June-2017.aspx], some rain gauges in the watershed indicated that this was the highest recorded 1-day total rainfall for that area since record-keeping began in 1950. The GRCA notes that the public ECCC weather forecast site predicted 5–10 mm of rainfall over the watershed area on 23 June 2017. But over a 2-day period, totals of precipitation were exceeding 100 mm and in some cases 130 mm at several stations in the watershed. The majority of rainfall occurred between 0700 and 1000 UTC 23 June based on ground rain gauge observations within the Grand River watershed.

The CaSPAr archive was already in place for one month in June 2017 (although data access was not possible until the automatic user request system went online in June 2018). The 7-day time delay of making data accessible would have not allowed the GRCA to draw the current forecasts. But the archive can help to show which data would have been available from ECCC and allows for a detailed postevent analysis comparing for example, the various ECCC precipitation

products. Therefore, we collected the CaSPAr data for Grand River watershed draining to gauge station in West Montrose, which was the watershed that was faced with those severe rainfall.

At the moment users can download operational, real-time weather forecast data via ECCC's Datamart portal. The data are available in grib2 and are interpolated to a grid the grib2 format supports. Accessing Datamart data for regular, gridded weather forecasts for a specific basin requires a notable effort and is not as straightforward as accessing CaSPAr data. Therefore, this sort of postevent analysis using CaSPAr data helps potential ECCC forecast users to evaluate if they should invest the effort to regularly access the operational data via Datamart. Spatial subsetting and the constraining of forecast horizons is not possible in Datamart. Hence, the subsetting of CaSPAr leads to significant reductions of data that need to be downloaded. In the case of the spatial subsetting of the Grand River watershed for this exercise, the reduction of data provided to the user was more than 99% for most of the products (GDPS: −99.91%; GEPS: −99.33%; RDPS: −99.89%; REPS: −99.02%; HRDPS: −99.96%; CaPA coarse: −96.52%; CaPA fine: −99.16%). The watershed extent and the grid cells downloaded from CaSPAr are shown in the maps in Fig. 3 (watershed extent as white outline and grid cells as gray tiles). HRDPS and CaPA (fine) have the highest resolution (2.5 km) and are precisely covering the domain while GEPS is very coarse (50 km) and hence only roughly approximates the 1,148-km^2 watershed. The number of grid cells with relevant data for the watershed is also listed in Table 1 [first line: Grand River at West Montrose (WM)].

Five of ECCC's NWPs contain precipitation and are available in June 2017—CaLDAS does not contain precipitation as it is meant to augment CaPA and RDRS is only available from 2010 to 2014. The five remaining products have different spatial resolutions and a different frequency of issue. GDPS is issued at midnight and noon (UTC) at a 25-km spatial resolution. GEPS is issued at same time as GDPS but contains 21 ensemble members and has a spatial resolution of 50 km. RDPS is issued every 6 h starting at midnight (UTC) and has a resolution of 10 km while REPS is issued only twice a day (midnight and noon UTC) for all 21 ensemble members at a resolution of 15 km. HRDPS has a resolution of 2.5 km and is issued four times a day starting at midnight (UTC). The CaPA products are available at 10 and 2.5 km (CaPA coarse and fine, respectively) and are both issued every 6 h starting at midnight (UTC). CaPA is an analysis, not a forecast, that assimilates observed rainfall of gauge stations and radars over the last 6 h and augments data-scarce regions with modeled estimates (Carrera et al. 2009; Lespinas et al. 2015; Fortin et al. 2018). The CaPA products (Figs. 3k–n) serve here as an observation rather than a forecast. The CaPA analysis at the fine (2.5 km) resolution has been added as a reference (gray bars) to the five available forecast products (Figs. 3b,d,f,h,j).

In the following we will analyze the (spatial) average of forecasted rainfall amounts over the entire domain of the Grand River watershed draining to West Montrose until midnight (UTC) on 24 June 2017. The rainfall rates (mm h^{-1}) are shown as lines in Fig. 3. The issue date of each time series is indicated as triangular markers of the same color below the corresponding start of the time series. The area under the lines corresponds to the average rainfall volume (mm) over the entire watershed.

The data show that indeed forecasts issued on 22 June 2017 were all forecasting a rainfall event but of lower volume. GDPS forecasted 40 mm until 0000 UTC 24 June in its midnight issue on 22 June (orange line in Fig. 3b) and only 20 mm in the noon (UTC) issue on 22 June (red line in Fig. 3b). The same holds for all forecasts issued on 22 June for GEPS, RDPS, and REPS. They all forecasted at most 40 mm of rain over the entire watershed until midnight (UTC) 24 June. The only exception is the high-resolution product HRDPS that forecasted, at 0600 UTC 22 June, 63.8 mm (light orange line in Fig. 3j) and, at 1200 UTC 22 June, 59.8 mm (red line in Fig. 3j) of precipitation until midnight (UTC) 24 June 2017.

The midnight (UTC) issue on 23 June 2017 showed high rainfall amounts in all deterministic forecasts. GDPS forecasted 57.3 mm, RDPS had 52.4 mm, and HRDPS forecasted 46.9 mm
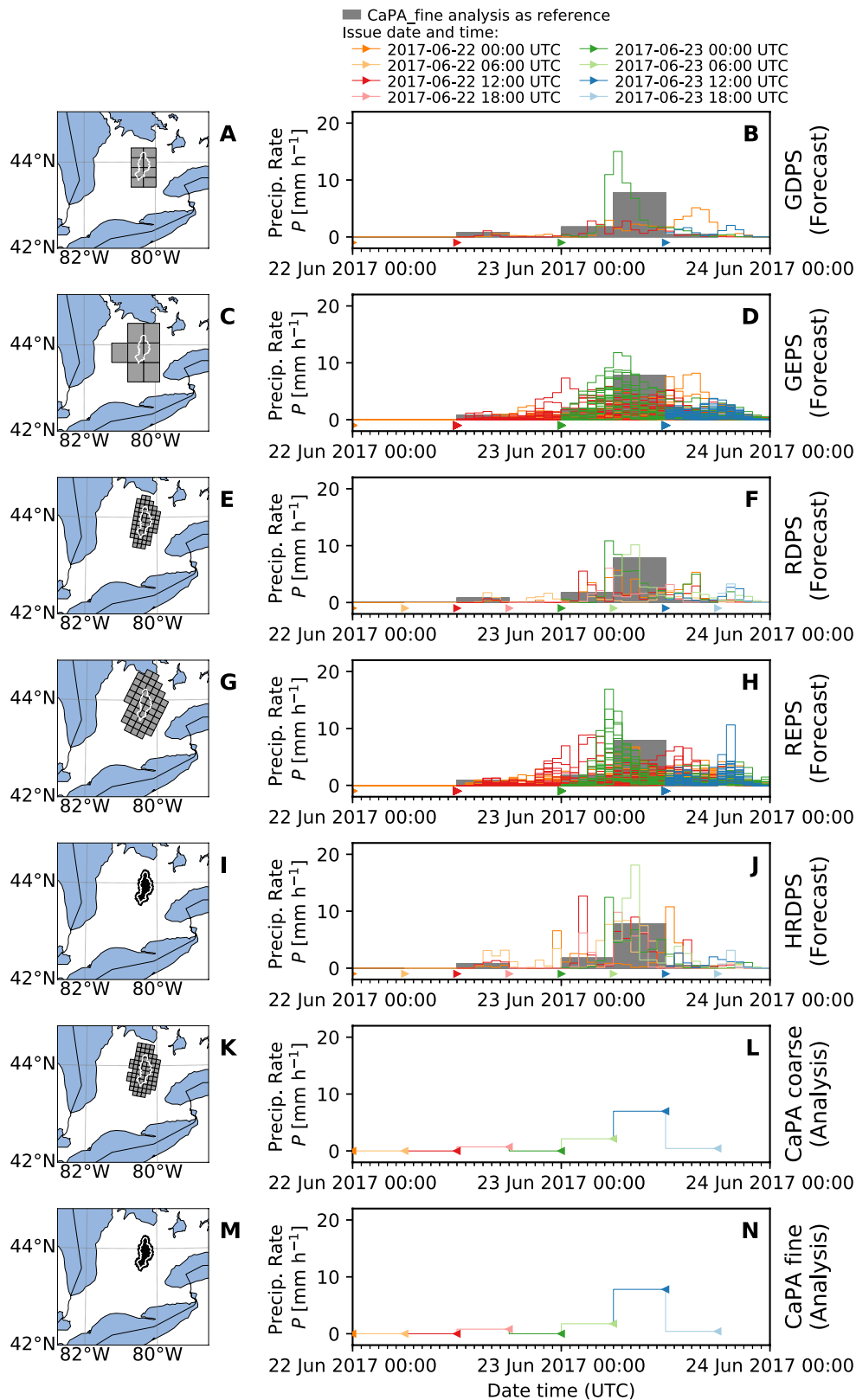
Fig. 3. Revisiting a record flood event that occurred at noon on 23 Jun 2017 (UTC) in the Grand River watershed close to West Montrose (highlighted with white outline in maps). (left) The grid cells covering this domain of interest (gray tiles) for the seven NWP products available in CaSPAr that contain precipitation and are available for this time period. RDRS was discarded since it is only available from 2010 to 2014. CaLDAS does not contain precipitation. (right) The total precipitation rate for the whole watershed for different forecasts and analyses (colored lines). The triangle markers indicate the issue date and time of each forecast; for example, GDPS is issued at noon (red and blue) and midnight (orange and green). The precipitation of the CaPA analysis at the fine resolution (2.5 km) serves as a ground truth and has been added for reference to the five forecast NWPs [gray bars in (b), (d), (f), (h), and (j)].

until midnight 24 June 2017. The ensemble forecasts were on average smaller but with large standard deviations between the individual forecasts. GEPS reported a mean of 34.8 mm (the 10th and 90th percentiles are p10 = 14.9 mm and p90 = 61.0 mm, respectively) and REPS forecasted on average 31.0 mm (p10 = 15.6 mm and p90 = 50.6 mm). The forecasts continued to be high in the 0600 UTC issue on 23 June 2017. The two available forecasts were RDPS with 40.9 mm and HRDPS with 50.2 mm within the next 18 h (until midnight of 24 June). After that all forecasts went back to lower volumes of about 10 mm or less. All of these forecasted amounts are not as high as the 100–130 mm reported by the Conservation Authority. This is at least partially due to the fact that with the West Montrose gauge drainage basin, we are comparing peak ground-based total event precipitation observations with average precipitation forecasts for the watershed. Continued analyses to reconcile ground-based observations with the various NWP products is also supported by CaSPAr and could focus on comparing the observations to corresponding gridcell values of the various NWP products.

Analyzing the rainfall of individual grid cells shows that HRDPS consistently forecasted grid cells with large rainfall amounts until midnight (UTC) 24 June 2017 (e.g., 0600 UTC 22 June: 91 mm; 1200 UTC 22 June: 86 mm; 1800 UTC 22 June: 62 mm; 0000 UTC 23 June: 60 mm; 0600 UTC 23 June: 81 mm). This holds also for the other four products. In all issues and ensemble members there were grid cells forecasted with extreme event totals more closely approaching the extreme ground-based observation totals for the event.

The comparison of the forecasts with the CaPA analysis of the fine resolution (gray bars added for reference to Figs. 3b, 3d, 3f, 3h, and 3j) clearly demonstrates that there is good consistency between some of the forecasts and CaPA (fine). Further comparative analyses between CaPA fine and the rainfall gauges and the forecasts at the measured rainfall gauges would be possible with CaSPAr and potentially of interest to the GRCA but these are not reported on here.

CaSPAr provides an opportunity for forecasting agencies such as the GRCA to test retrospectively alternative ECCC weather forecast products and understand the strengths and limitations of their current forecast systems and develop new forecast products. CaSPAr aims to improve the understanding and dissemination of ECCC's available NWPs. The above example shows that it is straight forward to compare several products due to the similar file structure. The data are all available at the same location and are not interpolated. The user only needs to download and process the data of interest. In the case of the West Montrose flooding event it is 12.9 MB of CaSPAr data compared to 4.9 GB of data the user would need to download from Datamart.

Conclusions and future work

The Canadian Surface Prediction Archive (CaSPAr) is an archive of Environment and Climate Change Canada's (ECCC's) numerical weather predictions (NWPs) and reanalysis products. The archive is free of charge and accessible to everybody. It serves as a platform to communicate and document ECCC products. It has been demonstrated that the standardized file format (CF-1.6-compliant netCDF-4 format) makes it easy to compare products. The file format can also directly or with minimal effort be ingested by a wide range of hydrologic and land surface models and modeling frameworks such as Raven (Craig 2019), Structure for Unifying Multiple Modeling Alternatives (SUMMA) (Clark et al. 2015a,b), variable infiltration capacity (VIC) (Liang et al. 1994), WRF-Hydro (Gochis et al. 2018), Noah-MP (Niu et al. 2011), Modélisation Environnementale Communautaire (MEC)—Surface and Hydrology (MESH) (Pietroniro et al. 2007), and mesoscale hydrologic model (MHM) (Samaniego et al. 2010). The subsetting functionality might also be helpful beyond hydrologic applications, for example, to setup atmospheric modeling experiments over limited (regional) areas. The download is handled via Globus (Foster 2011; Allen et al. 2012) for a fast and secure file

download free of charge for end users. The modular setup and separation of the front end and the back end of CaSPAr allows for the easy transfer of the framework to other systems, for example, to ECCC. A move of the system onto ECCC's servers is regarded to be beneficial since it would allow for a dissemination of the current forecasts without the 7-day time delay, which is only in place to make sure that file transfers from ECCC onto the current CaSPAr storage on Compute/Calcul Canada's system is successful and completed. The CaSPAr system fills the void of an accessible archive of the Canadian NWPs and brings them to the same level as NOAA's NOMADS system and the European NWPs disseminated by the ECMWF via the Climate Data Store. The CaSPAr system is the only one of these three systems that allows for spatial cropping and selection of forecasted horizons, issues, variables, and time periods. We hope that this will help practitioners such as conservation authorities to improve their flood forecasting and warning systems and help researchers to test and benchmark new data assimilation techniques and/or postprocessing algorithms (Han and Coulibaly 2019) by employing this archive of NWPs.

Ongoing developments of CaSPAr are targeting a full establishment of an application program interface (API) such that advanced users can avoid the web interface and can directly request data. We are also investigating the feasibility of implementing an OPeNDAP (www .opendap.org) access to the CaSPAr database.

Appendix: Details on products archived in CaSPAr.

The details on period of the archived data, forecasted horizons, number of issues per day, ensemble size, number of variables archived, and approximate spatial resolution are constantly updated on our GitHub wiki associated with the CaSPAr archive (https://github.com/julemai/CaSPAr/wiki/Available-products). We, however, added Table A1 for the convenience of the reader. Please note that all information given is subject to change since the operational products are constantly updated and improved by ECCC in the variables that are forecasted, and the spatial and temporal resolution. Currently 10.8 TB of data are added every month to CaSPAr.

Table A1. The products archived in CaSPAr so far are five operational forecasts (GEPS, GDPS, REPS, RDPS, and HRDPS), one reanalysis product (RDRS), three operational analyses (CaLDAS, CaPA coarse and fine), and one preoperational analysis (CaPA fine exp). The number of issues per day varies between the products. The first new estimate each day is issued midnight for all products. Three products are ensemble products (GEPS, REPS, and CaLDAS). The number of variables archived ranges between 2 and 55 while not all variables might yet be open for all users. The given spatial resolution is an average estimate per product. All the information given is subject to change as the operational products are constantly improved by ECCC. The total amount of new data stored every month is 11,025 GB (10.8 TB).

Product	Archived since	Archived until	Forecasted horizons (h)	Issues (day⁻¹)	Ensemble size	No. of variables	Spatial resolution (km)	Memory (GB month⁻¹)
GEPS	23 May 2017	Today	1, 2, …, 72, 96, …, 384	2	21	55	50.0	6,700
GDPS	22 May 2017	Today	1, 2, …, 144, 147, …, 240	2	1	55	25.0	506
REPS	23 May 2017	Today	1, 2, …, 72	2	21	52	15.0	1,902
RDPS	1 Jan 2015	Today	1, 2, …, 84	4	1	55	10.0	723
HRDPS	22 May 2017	Today	1, 2, …, 48	4	1	50	2.5	910
RDRS	1 Jan 2010	31 Dec 2014	1, 2, …, 24	1	1	12	15.0	5
CaLDAS	22 May 2017	Today	—	8	25	14	2.5	281
CaPA coarse	17 Sep 2012	Today	—	4	1	2	10.0	1
CaPA fine	3 Mar 2018	Today	—	4	1	2	2.5	2
CaPA fine exp	1 Jun 2016	2 Mar 2018	—	4	1	2	2.5	2

References

Abaza, M., F. Anctil, V. Fortin, and R. Turcotte, 2013: A comparison of the Canadian global and regional meteorological ensemble prediction systems for short-term hydrological forecasting. *Mon. Wea. Rev.*, **141**, 3462–3476, https://doi.org/10.1175/MWR-D-12-00206.1.

Allen, B., and Coauthors, 2012: Software as a service for data scientists. *Commun. ACM*, **55**, 81–88, https://doi.org/10.1145/2076450.2076468.

Balsamo, G., J.-F. Mahfouf, S. Bélair, and G. Deblonde, 2007: A land data assimilation system for soil moisture and temperature: An information content study. *J. Hydrometeor.*, **8**, 1225–1242, https://doi.org/10.1175/2007JHM819.1.

Cao, Z., and D.-L. Zhang, 2016: Analysis of missed summer severe rainfall forecasts. *Wea. Forecasting*, **31**, 433–450, https://doi.org/10.1175/WAF-D-15-0119.1.

Carrera, M. L., V. Fortin, and S. Bélair, 2009: Development of the Canadian Precipitation Analysis (CaPA) and the Canadian Land Data Assimilation System (CaLDAS). *DRI Precipitation and Drought Indices Workshop*, Toronto, ON, Canada, Environment Canada, www.meteo.mcgill.ca/dri/09precip/marco.pdf.

Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2009: Toward random sampling of model error in the Canadian ensemble prediction system. *Mon. Wea. Rev.*, **138**, 1877–1901, https://doi.org/10.1175/2009MWR3187.1.

Clark, M. P., and Coauthors, 2015a: A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resour. Res.*, **51**, 2498–2514, https://doi.org/10.1002/2015WR017198.

——, and Coauthors, 2015b: A unified approach for process-based hydrologic modeling: 2. Model implementation and case studies. *Water Resour. Res.*, **51**, 2515–2542, https://doi.org/10.1002/2015WR017200.

CMC, 2018a: High Resolution Deterministic Precipitation Analysis System (CaPA-HRDPA). Environment and Climate Change Canada Tech. Note, 16 pp., http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/tech_notes/technote_capa_hrdpa-450_e.pdf.

——, 2018b: High Resolution Deterministic Prediction System (HRDPS) version 5.0.0. Canadian Meteorological Centre Tech. Note, 57 pp., http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/tech_notes/technote_hrdps-500_20180918_e.pdf.

——, 2018c: Regional Deterministic Precipitation Analysis System (CaPA-RDPA). Environment and Climate Change Canada Tech. Note, 16 pp., http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/tech_notes/technote_capa_rdpa-450_e.pdf.

——, 2018d: Regional Ensemble Prediction System (REPS) version 2.4.0. Environment and Climate Change Canada Tech. Rep., 23 pp., http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/tech_notes/technote_reps-240_20180918_e.pdf.

——, 2018e: The Global Deterministic Prediction System (GDPS) version 6.1.0 of the Meteorological Service (MSC) of Canada. Environment and Climate Change Canada Tech. Rep., 9 pp., http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/tech_specifications/tech_specifications_GDPS_6.1.0_e.pdf.

——, 2018f: The Global Ensemble Prediction System (GEPS) version 5.0.0 of the Meteorological Service (MSC) of Canada. Environment and Climate Change Canada Tech. Rep., 10 pp., http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/tech_specifications/tech_specifications_GEPS_5.0.0_e.pdf.

——, 2018g: The Regional Deterministic Prediction System (RDPS) version 6.0.0 of the Meteorological Service of Canada (MSC). Environment and Climate Change Canada Tech. Rep., 7 pp., http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/tech_specifications/tech_specifications_RDPS_6.0.0_e.pdf.

Craig, J. R., 2019: Raven: User's and developer's manual v2.9.1. Raven Development Team Rep., 180 pp., http://raven.uwaterloo.ca/files/v2.9.1/RavenManual_v2.9.1.pdf.

Ficchì, A., L. Raso, D. Dorchies, F. Pianosi, P.-O. Malaterre, P.-J. V. Overloop, and M. Jay-Allemand, 2016: Optimal operation of the multireservoir system in the Seine River basin using deterministic and ensemble forecasts. *J. Water Resour. Plann. Manage.*, **142**, 05015005, https://doi.org/10.1061/(ASCE)WR.1943-5452.0000571.

Foley, A. M., P. G. Leahy, A. Marvuglia, and E. J. McKeogh, 2012: Current methods and advances in forecasting of wind power generation. *Renewable Energy*, **37**, 1–8, https://doi.org/10.1016/j.renene.2011.05.033.

Fortin, V., G. Roy, N. Donaldson, and A. Mahidjiba, 2015: Assimilation of radar quantitative precipitation estimations in the Canadian Precipitation Analysis (CaPA). *J. Hydrol.*, **531**, 296–307, https://doi.org/10.1016/j.jhydrol.2015.08.003.

——, ——, T. Stadnyk, K. Koenig, N. Gasset, and A. Mahidjiba, 2018: Ten years of science based on the Canadian precipitation analysis: A CaPA system overview and literature review. *Atmos.–Ocean*, **56**, 178–196, https://doi.org/10.1080/07055900.2018.1474728.

Foster, I., 2011: Globus online: Accelerating and democratizing science through cloud-based services. *IEEE Internet Comput.*, **15** (3), 70–73, https://doi.org/10.1109/MIC.2011.64.

Gagnon, N., and Coauthors, 2015: Global Ensemble Prediction System (GEPS)—Update from version 4.0.1 to version 4.1.1. Canadian Meteorological Centre Tech. Note, 36 pp.

Gochis, D. J., and Coauthors, 2018: The NCAR WRF-Hydro modeling system technical description (version 5.0). NCAR Tech. Rep., 107 pp.

Han, S., and P. Coulibaly, 2019: Probabilistic flood forecasting using hydrologic uncertainty processor with ensemble weather forecasts. *J. Hydrometeor.*, **20**, 1379–1398, https://doi.org/10.1175/JHM-D-18-0251.1.

Houtekamer, P. L., X. Deng, H. L. Mitchell, S.-J. Baek, and N. Gagnon, 2014: Higher resolution in an operational ensemble Kalman filter. *Mon. Wea. Rev.*, **142**, 1143–1162, https://doi.org/10.1175/MWR-D-13-00138.1.

Lavaysse, C., M. Carrera, S. Bélair, N. Gagnon, R. Frenette, M. Charron, and M. K. Yau, 2013: Impact of surface parameter uncertainties within the Canadian Regional Ensemble Prediction System. *Mon. Wea. Rev.*, **141**, 1506–1526, https://doi.org/10.1175/MWR-D-11-00354.1.

Lespinas, F., V. Fortin, G. Roy, P. Rasmussen, and T. Stadnyk, 2015: Performance evaluation of the Canadian Precipitation Analysis (CaPA). *J. Hydrometeor.*, **16**, 2045–2064, https://doi.org/10.1175/JHM-D-14-0191.1.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges, 1994: A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res.*, **99**, 14415–14428, https://doi.org/10.1029/94JD00483.

Lin, H., N. Gagnon, S. Beauregard, R. Muncaster, M. Markovic, B. Denis, and M. Charron, 2016: GEPS-based monthly prediction at the Canadian Meteorological Centre. *Mon. Wea. Rev.*, **144**, 4867–4883, https://doi.org/10.1175/MWR-D-16-0138.1.

Mahfouf, J.-F., B. Brasnett, and S. Gagnon, 2007: A Canadian Precipitation Analysis (CaPA) project: Description and preliminary results. *Atmos.–Ocean*, **45**, 1–17, https://doi.org/10.3137/ao.v450101.

Milbrandt, J. A., S. Bélair, M. Faucher, M. Vallée, M. L. Carrera, and A. Glazer, 2016: The pan-Canadian high resolution (2.5 km) deterministic prediction system. *Wea. Forecasting*, **31**, 1791–1816, https://doi.org/10.1175/WAF-D-16-0035.1.

Niu, G. Y., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res.*, **116**, D12109, https://doi.org/10.1029/2010JD015139.

Pietroniro, A., and Coauthors, 2007: Development of the MESH modelling system for hydrological ensemble forecasting of the Laurentian Great Lakes at the regional scale. *Hydrol. Earth Syst. Sci.*, **11**, 1279–1294, https://doi.org/10.5194/hess-11-1279-2007.

Ren, Y., P. Suganthan, and N. Srikanth, 2015: Ensemble methods for wind and solar power forecasting——A state-of-the-art review. *Renewable Sustainable Energy Rev.*, **50**, 82–91, https://doi.org/10.1016/j.rser.2015.04.081.

Rogelis, M. C., and M. Werner, 2018: Streamflow forecasts from WRF precipitation for flood early warning in mountain tropical areas. *Hydrol. Earth Syst. Sci.*, **22**, 853–870, https://doi.org/10.5194/hess-22-853-2018.

Rutledge, G. K., J. Alpert, and W. Ebisuzaki, 2006: NOMADS: A climate and weather model archive at the National Oceanic and Atmospheric Administration. *Bull. Amer. Meteor. Soc.*, **87**, 327–342, https://doi.org/10.1175/BAMS-87-3 -327.

Samaniego, L., R. Kumar, and S. Attinger, 2010: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resour. Res.*, **46**, W05523, https://doi.org/10.1029/2008WR007327.

Schwanenberg, D., F. M. Fan, S. Naumann, J. I. Kuwajima, R. A. Montero, and A. Assis dos Reis, 2015: Short-term reservoir optimization for flood mitigation under meteorological and hydrological forecast uncertainty. *Water Resour. Manage.*, **29**, 1635–1651, https://doi.org/10.1007/s11269 -014-0899-1.

Soman, S. S., H. Zareipour, O. Malik, and P. Mandal, 2010: A review of wind power and wind speed forecasting methods with different time horizons. *North American Power Symp. 2010*, Arlington, TX, IEEE, https://doi.org/10.1109 /NAPS.2010.5619586.

Taylor, J. W., and R. Buizza, 2003: Using weather ensemble predictions in electricity demand forecasting. *Int. J. Forecasting*, **19**, 57–70, https://doi.org/10.1016 /S0169-2070(01)00123-6.

Thiboult, A., F. Anctil, and M. H. Ramos, 2017: How does the quantification of uncertainties affect the quality and value of flood early warning systems? *J. Hydrol.*, **551**, 365–373, https://doi.org/10.1016/j.jhydrol.2017.05.014.

Xuan, Y., I. D. Cluckie, and Y. Wang, 2009: Uncertainty analysis of hydrological ensemble forecasts in a distributed model utilising short-range rainfall prediction. *Hydrol. Earth Syst. Sci.*, **13**, 293–303, https://doi.org/10.5194/hess-13-293-2009.

Yang, S.-C., and T.-H. Yang, 2014: Uncertainty assessment: Reservoir inflow forecasting with ensemble precipitation forecasts and HEC-HMS. *Adv. Meteor.*, **2014**, 581756, https://doi.org/10.1155/2014/581756.

Yucel, I., A. Onen, K. Yilmaz, and D. Gochis, 2015: Calibration and evaluation of a flood forecasting system: Utility of numerical weather prediction model, data assimilation and satellite-based rainfall. *J. Hydrol.*, **523**, 49–66, https://doi .org/10.1016/j.jhydrol.2015.01.042.