

A Framework for Joint Verification and Evaluation of Seasonal Climate Services across Socioeconomic Sectors

Louise Crochemore^{1a,o} Stefano Materia^{b,p} Elisa Delpiazzi^{c,d,e} Stefano Bagli^f
Andrea Borrelli^b Francesco Bosello^{c,d,e} Eva Contreras^g Francesco Dalla Valleⁱ
Silvio Gualdi^b Javier Herrero^{g,r} Francesca Larosa^{j,k,m} Rafael Lopez^l Valerio Luzzi^f
Paolo Mazzoli^f Andrea Montani^{h,n} Isabel Moreno^l Valentina Pavan^h
Ilias Pechlivanidis^a Fausto Tomei^h Giulia Villani^h Christiana Photiadou^{a,q}
María José Polo^g and Jaroslav Mysiak^{j,k}

KEYWORDS:

Forecast
verification/skill;
Seasonal
forecasting;
Climate services;
Economic value

ABSTRACT: Assessing the information provided by coproduced climate services is a timely challenge, given the continuously evolving scientific knowledge and its increasing translation to address societal needs. Here, we propose a joint evaluation and verification framework to assess prototype services that provide seasonal forecast information based on the experience from the Horizon 2020 (H2020) Climate forecasts enabled knowledge services (CLARA) project. The quality and value of the forecasts generated by CLARA services were first assessed for five climate services utilizing the Copernicus Climate Change Service seasonal forecasts and responding to knowledge needs from the water resources management, agriculture, and energy production sectors. This joint forecast verification and service evaluation highlights various skills and values across physical variables, services, and sectors, as well as a need to bridge the gap between verification and user-oriented evaluation. We provide lessons learned based on the service developers' and users' experience and recommendations to consortia that may want to deploy such verification and evaluation exercises. Last, we formalize a framework for joint verification and evaluation in service development, following a transdisciplinary (from data purveyors to service users) and interdisciplinary chain (climate, hydrology, economics, and decision analysis).

SIGNIFICANCE STATEMENT: Tools to communicate climate-related information to users, typically dam managers, irrigation consortia, or energy producers, are fast evolving to answer societal needs. It is crucial to estimate the quality of the provided information, along with economic, environmental, and/or societal gains. Here, we exemplify how to assess information quality and potential gains in five services that provide data and information for hydropower, solar power, irrigation, and water reservoirs in Europe and South America. Based on this work, we recommend 1) service developers to well anticipate such quality and value assessments, due to the number of actors to be involved; 2) flexibility when screening how to quantify quality and gain to account for decision contexts; and 3) sustained funding or collaborating platforms to ensure the iterative coevaluation process.

DOI: 10.1175/BAMS-D-23-0026.1

Corresponding author: Louise Crochemore, louise.crochemore@univ-grenoble-alpes.fr

Manuscript received 10 February 2023, in final form 12 March 2024, accepted 16 March 2024

© 2024 American Meteorological Society. This published article is licensed under the terms of the default AMS reuse license. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

AFFILIATIONS: ^a Swedish Meteorological and Hydrological Institute, Hydrology Research Unit, Norrköping, Sweden; ^b CSP Division, Centro Euro-Mediterraneo sui Cambiamenti Climatici, Bologna, Italy; ^c ECIP Division, Centro Euro-Mediterraneo sui Cambiamenti Climatici, Marghera, Italy; ^d Department of Environmental Sciences, Informatics and Statistics Department, Ca' Foscari University of Venice, Mestre, Italy; ^e European Institute for the Economy and Environment, Marghera, Italy; ^f GECOSistema Srl, Cesena, Italy; ^g Fluvial Dynamics and Hydrology Research Group, Andalusian Institute for Earth System Research, University of Córdoba, Córdoba, Spain; ^h Arpa^e—Regional Agency for Prevention, Environment and Energy of Emilia-Romagna, Bologna, Italy; ⁱ Enel Green Power, MTS—Dams and Civil Infrastructures Safety—Hydrology and Hydraulic Analysis, Mestre, Italy; ^j RAAS Division, Centro Euro-Mediterraneo sui Cambiamenti Climatici, Marghera, Italy; ^k Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Mestre, Italy; ^l Physics for Renewable Energies Research Group, Universidad de Córdoba, Córdoba, Spain; ^m KTH Climate Action Centre, Royal Institute of Technology, Stockholm, Sweden; ⁿ European Center for Medium-Range Weather Forecasts, Reading, United Kingdom; ^o Université Grenoble Alpes, CNRS, INRAE, IRD, Grenoble INP, IGE, Grenoble, France; ^p Barcelona Supercomputing Center, Barcelona, Spain; ^q European Environment Agency, Copenhagen, Denmark; ^r Global Ecology Research Group/Research Unit Modeling Nature, Faculty of Sciences, University of Granada, Granada, Spain

1. Introduction

Climate and water services convey climate-relevant knowledge and information (processed, structured, and well-communicated data interpreted for decision-making contexts) for horizons ranging from weeks to decades ahead. Climate services, hereafter used to refer to both climate and water services, have been recognized as pivotal instruments to support decision-making at all levels, from public entities to business operators (Alexander and Dessai 2019; Boon et al. 2022; Brasseur and Gallardo 2016) and across sectors, such as disaster risk reduction, energy production, water resources management, or agriculture (e.g., Street et al. 2019; Troccoli 2018).

One key conclusion that emerged from previous service generation initiatives is the prominent role of users to ensure actionable information, and thus the need for coproduction that ensures continuous interactions between service users, developers, and data purveyors (Bremer et al. 2019; Photiadou et al. 2021; Cantone et al. 2023). Codevelopment and coevaluation are the participatory steps of coproduction (together with codesign and co-delivery; Mauser et al. 2013) which focus on developing a climate information chain addressing user needs and on assessing the value of climate services, respectively. By building trust through direct participation, codevelopment and coevaluation promote the sustainable use of climate services and contribute to the common reflection on how climate data may address user needs (Clements et al. 2013).

The need for “standards for what constitutes ‘quality’ climate information” and improved skill indicators to support users with measurable quality criteria has been previously highlighted (Vaughan and Dessai 2014). Forecast verification, which assesses the quality of forecasts in reproducing retrospective outcomes (Jolliffe and Stephenson 2003; Wilks 2011), is somewhat common practice, though not systematic, in openly available climate services. Yet, forecast verification does not account for the decision being informed by forecasts, and therefore does not capture the benefit (utility) and economic value of the forecast information (Giuliani et al. 2020). The literature on socioeconomic forecast evaluation has thus grown, bridging the gap between the climate and economical science communities (e.g., Katz and

Lazo 2011; Laugesen et al. 2023). Nevertheless, examples of the value (economic, societal, or environmental) attained by climate services in use and how these assessments are practically carried out are far less common (Vaughan et al. 2019).

This paper aims, in a first stage, to present the skill and value assessments of five service prototypes based on seasonal forecasts (section 2). These services were developed within the Horizon 2020 innovation action Climate forecasts enabled knowledge services (CLARA) project (<https://www.clara-project.eu>, 2017–20). The project aimed to support the coproduction of climate service prototypes across sectors and demonstrate the implementation of the open-access and large-scale Copernicus Climate Change Service (C3S). It involved service developers, sectorial users, hydroclimate scientists, and economists in coproducing climate services and in ensuring their longevity through sustainable business models. Quantitative outcomes from the verification and evaluation exercises performed within CLARA are presented (sections 3 and 4) and qualitative outcomes from developers' and users' viewpoints are discussed (section 5). We draw recommendations for verification and evaluation exercises based on the CLARA experience (section 6) and formalize a framework for assessing the quality and value of climate services providing information at the seasonal time scale and focusing on operational decision-making (section 7), before concluding (section 8).

2. Design of the verification and evaluation exercises within CLARA

a. Climate services. The services studied here were selected among the 14 coproduced within CLARA because they make use of the seasonal forecasts available from C3S (rather than decadal information or climate projections). They aim to support decision-making through interactive forecast visualization and provide sector-oriented forecasts targeting water resource management, agriculture, hydropower production and solar energy production, and potential users such as water authorities, reservoir managers, and energy producers. Table 1 presents the objectives, developers, and users of each service, and Fig. 1 shows the data and information flowchart of each service.

The way seasonal forecasts are generated in each service depends on the specificities of each natural system and on the expertise of the service providers (Fig. 1 and Table 2a). For instance, the Smart Climate Hydropower Tool (SCHT) service was developed using artificial intelligence (AI) to relate meteorological predictands with reservoir inflows, and Solar Energy Assessment and Planning Tool (SEAP) is based on a statistical model that predicts photovoltaic plant production based on solar, wind, and meteorological conditions. Reservoir Operation Assessment Tool (ROAT) and Small Hydropower Management and Assessment Tool (SHYMAT) offer tailored information at local scale through bias adjustment and downscaling of available C3S seasonal forecasts. Water resources for irrigation (WRI) are based on a mechanistic impact modeling meant to translate hydrometeorological information into variables of interest for the agriculture sector. The concepts behind these services are transposable to different geographical areas. Nevertheless, the verification and evaluation presented hereafter focus on their original target areas in Europe (i.e., Spain and Italy) and South America (i.e., Colombia).

All services rely on multiple data sources, including observations acquired locally in real time or simulations from impact models and meteorological and hydrological forecasts from C3S available through the Climate Data Store (CDS) (see details in Table 2a): the fifth generation seasonal forecasts (SEAS5) from the European Centre for Medium-Range Weather Forecasts (ECMWF; Johnson et al. 2019), Seasonal Prediction System 3 (SPS3) from the Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC; Sanna et al. 2016), System 7 from Météo-France (Batté et al. 2019), and Global Framework for Climate Services (GFCS) 2.0 from DWD (Baehr et al. 2015). Seasonal hydrological forecasts were made available to service developers from the C3S proof-of-concept service for the European water resources

TABLE 1. General information on the CSs detailing their sector of application, their name and a corresponding reference, their aim, target audience, developer, user within CLARA, and a link for access.

Sector	Service name (developer)	Aim	Target audience	Developer	User involved in the codevelopment	Access
Water resources management	ROAT— <i>Reservoir Operation Assessment Tool</i>	Support operations of multiobjective reservoirs by anticipating (i) drought risks and setting a “scarcity level” and (ii) water excess from snowmelt to avoid damages downstream the dam.	Water authorities, reservoir managers	University of Córdoba	Technicians of the Béznar-Rules reservoir	https://www.uco.es/dfh/ROAT/
Agriculture	WRI— <i>Water Resources for Irrigation</i> (Villani et al. 2021)	Support decision-making for both water procurement and water allocation by anticipating water irrigation needs to increase efficiency and reduce irrigation water and energy consumption	Irrigation water management authorities (e.g., water procurement and allocation agencies)	Arpea	Land Reclamation and Irrigation Consortium of Romagna and Burana	https://servizi-gis.arpae.it/moses/home/index.html
Hydropower production	SCHT— <i>Smart Climate Hydropower Tool</i> (Essenfelder et al. 2020)	Simplify decision-making processes by predicting hydropower production for management and financial planning	Hydropower energy producers	Geographic Environmental Consulting (GECOSistema)	Enel Green Power	https://gecosistema.com/scht
	SHYMAT— <i>Small Hydropower Management and Assessment Tool</i> (Contreras et al. 2020a,b)	Support operation of run-of-river plants such as (i) planning the operation and maintenance tasks, (ii) anticipating energy production, and (iii) setting the turbine level depending on river flow	Technicians in charge of the run-of-river plants and managers of hydropower companies. Energy market operators, river basin authorities, and consultants	University of Córdoba	Endesa technicians	https://www.uco.es/dfh/SHYMAT/
Solar energy production	SEAP— <i>Solar Energy Assessment and Planning Tool</i>	Automate both spatial and operational assessments of utility-scale photovoltaic power plants by providing information about the optimal tracking system for dynamic collectors.	Photovoltaic plant managers	University of Córdoba	Magtel technicians	https://www.uco.es/investigacion/proyectos/seap/

management sector (<https://climate.copernicus.eu/operational-service-water-sector>). These forecasts are based on the European setup for Hydrological Predictions for the Environment (E-HYPE) hydrological model from the Swedish Meteorological and Hydrological Institute (SMHI) and the ECMWF SEAS5 meteorological forecasts (Pechlivanidis et al. 2020).

b. Methodology. Due to the transdisciplinary and time-limited nature of the exercise, compromises were made in selecting verification and evaluation methodologies. The metrics used to convey the forecast quality and service value were chosen so as to be easily computable by the service developers within the time of the project and easily understandable and translatable to users. This led, in some cases, to the simplification of known forecast attributes (reliability) and in the selection by users of reduced but contrasting sets of situations and actions. Neither the verification nor the evaluation metrics used here are part of the recommendations. As discussed thereafter, verification metrics should be tailored for each climate service and decision context, highlighting forecast attributes relevant to the evaluation.

Hydrometeorological observations were used as *reference* (or truth) in the verification and evaluation of the services, i.e., river flow observations (Table 2b; SHYMAT, SCHAT, and ROAT).

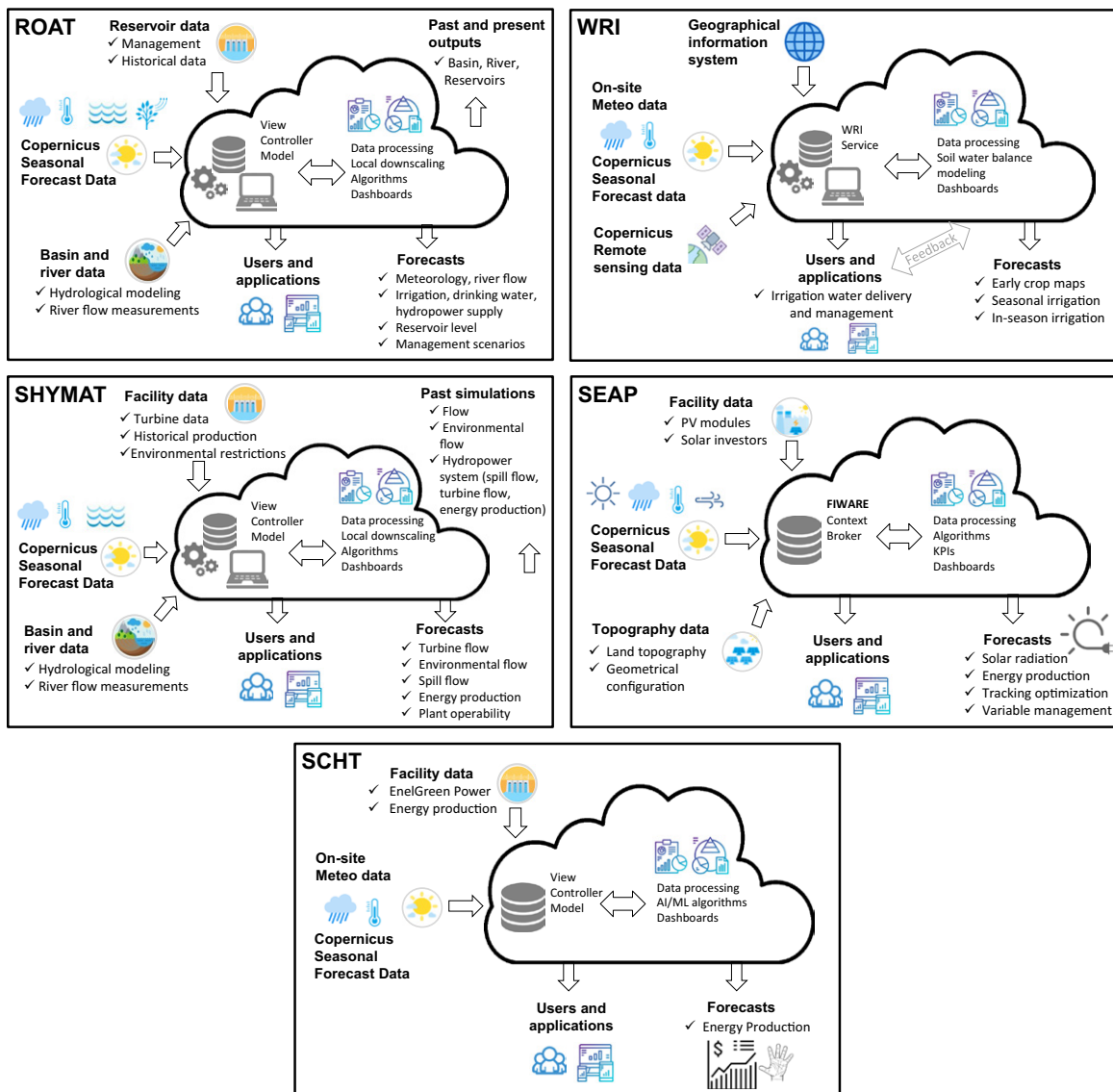


FIG. 1. Workflow diagrams of the CSs. Each service workflow displays the service input data, including the seasonal forecast data, the methods applied within the service (such as preprocessing, impact models, and user interface), and the type of output data. A detailed description of the verified and evaluated components is available in Table 2.

When local observations were not available, the fifth major global reanalysis produced by ECMWF (ERA5)-Land meteorological reanalysis data (Hersbach et al. 2020) forced the impact models used in the services to generate reference simulations (e.g., temperature and precipitation in the case of WRI to simulate crop water needs; solar radiation, temperature, and wind in the case of SEAP to simulate photovoltaic production).

1) DESIGNING THE VERIFICATION EXERCISE. Verification was performed for each service by comparing hindcasts (retrospective forecasts) of the service with a reference (or truth) over a long enough time period with respect to different *forecast attributes*. The relative quality of the service with respect to a reference forecast (or *benchmark*) was then assessed with *skill scores* (see Table 2b). The assessment was tailored to the specificities of each service, meaning that the seasonal forecast set, target variable, geographical area, and verified period and forecast attributes vary with the service based on user needs and data availability (Table 2b). Note that verification should ideally be performed with at least 30 years of hindcasts. The short verification periods for WRI and SEAP thus lead to significant uncertainties in their verification thereafter.

TABLE 2. CLARA services selected for the exercise: (a) characteristics and parameters for the (b) verification and (c) evaluation exercises.

(a)							
Service	ROAT	WRI		SCHT		SHYMAT	SEAP
Area considered	Órgiva River basin in Sierra Nevada (495 km ²), Spain	Irrigated area of the Irrigation Consortium of Romagna (930 km ²), Italy	Irrigated area of the Irrigation Consortium of Burana (600 km ²), Italy	Betania hydropower plant, Magdalena River (13 000 km ²), Colombia	Guavio hydropower plant on the Guavio River (1500 km ²), Colombia	Water-contributing area to Pampaneira hydropower plant (82.5 km ²)	“El Molino,” solar installation, located in Cordoba, Spain
Service output forecasts	Dynamical probabilistic forecasts bias adjusted with quantile mapping	Hybrid probabilistic forecasts from a weather generator supplying precipitation and temperature		AI-based statistical deterministic forecasts		Dynamical probabilistic forecasts downscaled for the system location	Statistical deterministic forecasts
Copernicus input forecasts	Forecasts of river flow as inflows to the reservoir (source: ECMWF/SMHI)	Forecasts of precipitation and temperature (source: CMCC/ECMWF /DWD/ MétéoFrance–100 members)		Forecasts of precipitation and temperature (source: CMCC)		Forecasts of river flow as inflows to the hydropower plant (source: ECMWF/SMHI)	Forecasts of wind, radiation, precipitation, and temperature (source: CMCC)
(b)							
Service	ROAT	WRI		SCHT		SHYMAT	SEAP
Quality-assessed variable	Bias-adjusted river flow (m ³ s ^{−1})	Crop irrigation water needs (m ³ ha ^{−1})		Inflow to the reservoirs (mm ³ month ^{−1})		Bias-adjusted river flow (m ³ s ^{−1})	Photovoltaic production (kWh month ^{−1})
Assessment period	22 years (January 1989–December 2014 excluding 1990, 1997–99)	6 years (2011–16)		26.5 years (January 1993–August 2019)		35 years (January 1981–November 2015)	10 years (January 1997–December 2006)
Time step	Monthly	3-month		6-month	3-month	Monthly	Monthly
Issue dates and horizons	Monthly for each of the 7 months ahead—12 × 7 forecast values per year	Once a year for the summed volume over June, July, and August—1 forecast value per year		Monthly for the inflows averaged over the next 6 months—12 forecast values per year	Monthly for the inflows averaged over the next 3 months—12 forecast values per year	Monthly for each of the 7 months ahead—12 × 7 forecast values per year	Monthly for the month ahead—12 forecast values per year
Reference—truth	River flow observations from Hidrosur and Béznar-Rules Reservoir System	Crop water needs reanalysis data based on ERA5-Land [one grid cell (0.1° × 0.1°)]		Inflow observations from Enel Green Power		River flow observations from Endesa	Photovoltaic production reanalysis data based on ERA5
Benchmark—reference forecast	Ensemble forecast based on historical flows from the previous 10 years	Forecast based on ERA5-Land climatology (1993–2010) as input to the model		Enel linear model fed with Copernicus seasonal forecasts (CMCC)		Ensemble forecast based on historical flows from the previous 10 years	Forecast based on ERA5 reanalysis as input to the photovoltaic production model
(c)							
Service	ROAT	SCHT		SHYMAT		SEAP	
Variable used to define the states of the world (indicator)	Stored water volumes (hm ³)—derived from river flow and reservoir rules	Stored water volume (mm ³)—derived from reservoir inflows and reservoir rules		River flow (m ³ s ^{−1})		Daily irradiation on photovoltaic modules based on clearness index, kt (kt values from 0 to 0.88)—correlated with photovoltaic production	
Test period	From 2010 to 2015	From 2011 to 2014		1-yr (2010/11) and 4-yr (2011–15) periods		2013	

(Continued)

Table 2. (Continued).

(c)				
Service	ROAT	SCHT	SHYMAT	SEAP
Time scale	Monthly	Monthly	Monthly	Daily
States of the world	Five states: (i) flood control, (ii) normal state, (iii) early warning, (iv) alert, and (v) emergency	Three states: water volume (i) inside, (ii) below, and (iii) above the normal range defined by the management rule in the previous 10 years	Four states: flow (i) higher than maximum flow, (ii) inside maximum range, (iii) inside minimum range, and (iv) lower than minimum flow	Eight states based on eight different levels of kt
Alternative knowledge source	Historical data of stored water volume and inflows	Climatological mean of water volumes over the last 30 years	Experience-based knowledge on river flow thresholds	Actual collectors' tracking method
Payoff	Payoff based on availability of water for different uses	Payoff based on produced energy	Payoff based on produced energy	Payoff based on solar radiation (proxy for profits)
Unit of measure for service value	Percent variation in performance index based on final users' judgments (0–10)	Percent variation in performance index based on final users' judgments (0–10)	Percent variation in performance index based on final users' judgments (0–10)	kWh m ⁻²

Even though probabilistic forecasts are recommended to convey the known uncertainties at seasonal lead times, not all decision frameworks incorporate probabilistic information yet. Some services thus provided deterministic forecasts as a support to current decision flows. For the services involving probabilistic forecasts (ROAT and SHYMAT), an indicator of *reliability* was complemented with an indicator of *sharpness* (often assessed alongside reliability; Gneiting et al. 2007). For the services based on deterministic forecasts (SCHT and SEAP), *correlation* and *accuracy* were assessed. In the case of WRI, an event-based verification assesses accuracy in identifying above- and below-average situations. All attributes were assessed for each initialization month and forecast month relevant to the service, using cross-validation when parameters were adjusted (see appendix A for metrics formulations). Skill scores assess the relative quality with respect to a forecast benchmark (Hargreaves 2010; appendix A), which consists of forecasts that decision-makers would use in the absence of the service (Table 2b). The benchmark choice was tailored to each service, with the goal to represent a realistic alternative to the proposed service. For four services, forecasts were compared to a climatological benchmark: ROAT and SHYMAT verified their forecasts against observed climatology, while WRI and SEAP used model climatology (i.e., simulations generated by forcing impact models with historically observed meteorological fields) due to the lack of observations of crop water needs and photovoltaic production. For SCHT, the forecasts are AI-based and were compared to a regression model also forced with Copernicus meteorological forecasts. Skill scores were computed for each forecast attribute, with positive values indicating that the service outperforms the benchmark.

2) DESIGNING THE EVALUATION EXERCISE. Service *value* (Table 2c) stems from the benefits from forecast-based decisions. Coevaluation can be *ex ante* (during codevelopment) demonstrating potential values, or *ex post* (after codevelopment) demonstrating real-life service value after co-delivery. Here, the *ex ante* service value was estimated for all services but WRI due to limited in situ observations related to irrigation practices.

Many reviews on climate service evaluation methods (e.g., Vaughan et al. 2019; Clements et al. 2013) have shown the impossibility to define a one-fits-all approach. The methodology

chosen here is a Bayesian probabilistic framework (Murphy 1993; Katz and Murphy 1997; Katz and Lazo 2011) with the advantage to produce a quantitative assessment fostering the climate service adoption, especially if the user is private and profit-seeking. The challenge of this fairly simple methodology resides in harmonizing inputs from different climate services and decision-making contexts, so that they are usable for the methodology and meaningful for the users (appendix B describes how it was practically performed). In CLARA, this approach was complemented by a qualitative evaluation (not presented here) to assess users' perception of the services beyond just profits. That (ex post) evaluation could, however, only be performed once the services were developed.

The link between the forecasts and the operative decisions is represented through an *indicator* considered by the user when making decisions. It could be the variable forecast by the service or a derived variable. Some critical indicator *thresholds* delimit intervals named *states of the world*. The users provide information about the *payoffs* they can gain when they decide to act based on a state of the world. An important aspect of the assessment is the metrics used to represent payoffs. A monetary or physical metric is easily understandable by users and the general public, though sometimes not possible because the service has not yet been used in a real-life decision-making context or the payoff information is sensitive. Alternatively, performance can be measured by asking decision-makers to assign scores to combinations of forecast states and actions (see "Unit of measure" in Table 2c).

The *expected value* is assessed for different knowledge sources based on the payoff of actions in each state of the world, the capacity of the knowledge source to forecast this state of the world, and on the probability of occurrence of this state of the world (for a comprehensive review and a practical example see Delpiazzi et al. 2022). First, the expected value of *perfect information* (Pope et al. 2017) is derived from the effective realizations (based on reference observations, or simulations) representing a climate service that perfectly forecasts the states of the world. It is instrumental as it demonstrates the maximum potential value of a service. The expected value of an *alternative knowledge* source (based on business-as-usual practices) is then assessed. In the case of SCHAT and ROAT, the alternative knowledge is climatology-based; for SHYMAT, it is the experience-based performance (replicating the same actions of a past period); and in SEAP, the operators' original tracking methods was used as reference. Last, once available, the expected value of the *climate service knowledge* is assessed.

The outcomes of the value assessment are the *effective value of the service* and its *maximum potential value*. The former is the difference in expected value between the climate service in its given development phase and the alternative knowledge (Table 2c). The latter derives from the difference between the expected value of perfect information when past observations are used as proxies for a climate service that always forecasts correctly and the alternative knowledge.

3. Outcomes of forecast verification

When verifying, deterministic forecasts of water volumes (SCHAT) and photovoltaic production (SEAP) show the highest accuracy, while probabilistic forecasts of crop water needs (WRI) show the lowest accuracy (not shown). Such results could be expected due to the high predictability of river flows as opposed to the low predictability of precipitation, that weighs directly on the performance of the water needs forecasts. In fact, the physiographic properties of the river systems act as filters to the highly variable precipitation input (Sutanto et al. 2020; Crochemore et al. 2020). Despite their low accuracy, WRI forecasts of irrigation needs seem to discriminate between below- and above-average conditions. Conversely, despite a seemingly high accuracy, forecasts of photovoltaic production reach low correlation levels, potentially due to irradiance being strongly dependent on cloud cover, which is unpredictable on time scales longer than a few days. Extending the verification periods used for WRI and SEAP

would, however, be necessary to confirm these performances. Last, both services applying the probabilistic verification framework reach reliability levels greater than 60% and can thus be considered acceptable to inform decision-making [cf. the cross-sector European-wide survey by Bruno-Soares et al. (2018) and the cross-sector worldwide serious game by Crochemore et al. (2021)].

The skill analysis (i.e., quality compared to that of benchmarks) provides a different perspective on the information provided by the services (Fig. 2). An increase in the quality of probabilistic forecasts with respect to the benchmarks requires an increase in sharpness without sacrificing reliability (Gneiting et al. 2007). Indeed, dynamical forecasts offer scenarios updated based on the latest ocean and atmosphere states and thus provide sharper ensembles than climatology-based forecasts, but sometimes at the cost of their reliability. Climatology-based ensembles offer perfect reliability, by nature, which services can, at best, equal, as in the case of SHYMAT (Fig. 2b). In the case of ROAT, where ensemble forecasts are sharper than climatology-based forecasts but less reliable, a trade-off between reliability and sharpness exists. This trade-off is well known and Weisheimer and Palmer (2014) have argued that “goodness” should be first assessed in terms of reliability (or calibration) of the ensemble forecasts, since only reliable inputs may be considered for any forecast-based decision-making.

The skill from deterministic forecasts varies with the service (Fig. 2a). SCHAT improves upon the linear regression in terms of river flow correlation, which can be attributed to the AI-based forecasting method. Forecasts of photovoltaic production (SEAP) also show improved

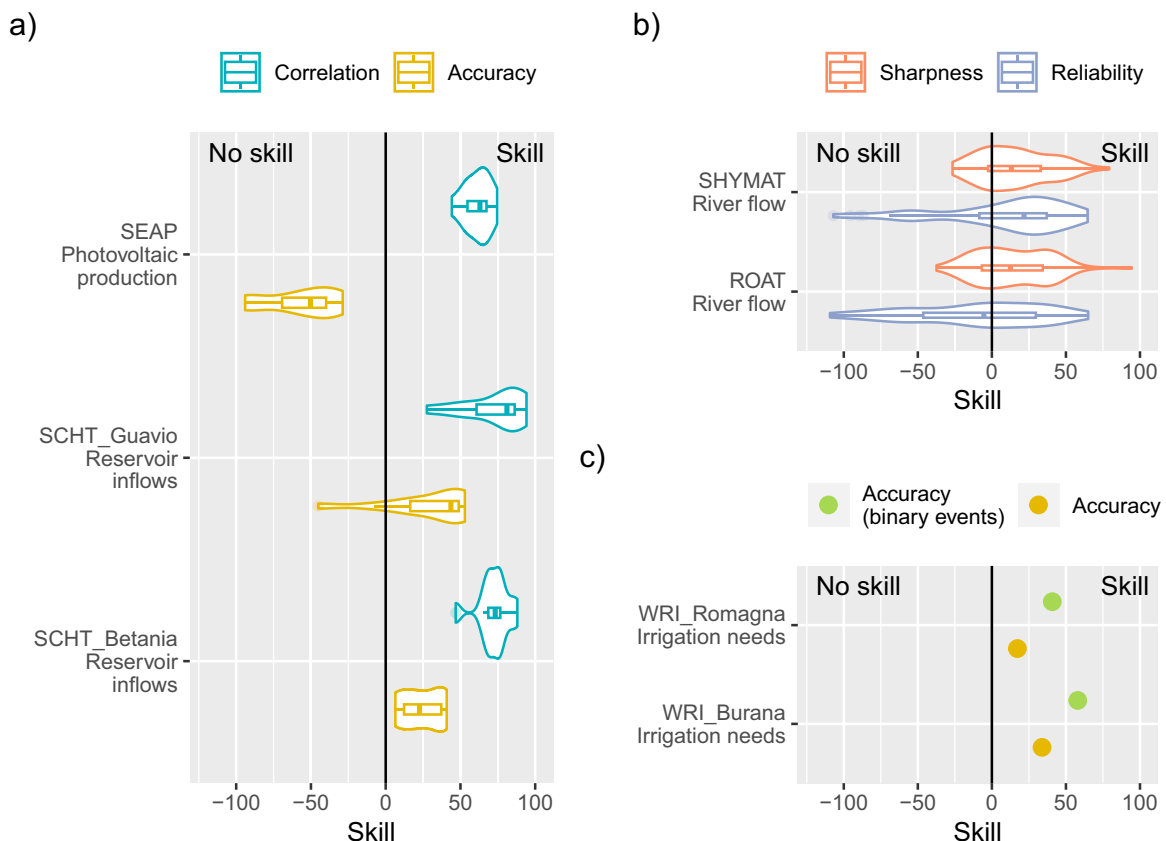


FIG. 2. Skill assessment of the services based on (a) deterministic and (b) probabilistic forecasts and (c) forecasts targeting threshold detection. The forecasts of the SCHAT and WRI services are verified for two locations: the Guavio and Betania hydropower plants for SCHAT, and Romagna and Burana for WRI. The black line indicates the skill value for which the service and benchmark forecasts have equivalent performances. Violins (showing the density) and boxplots (showing quartiles) indicate the variability in performance when the service supplies information at various time horizons (SHYMAT and ROAT) and/or initialization months (SEAP, SCHAT, SHYMAT, and ROAT). See Table 2b “Issue dates and horizons” for more details.

correlation compared to the benchmark despite low quality levels. In the case of SEAP, the high forecast accuracy does not translate in skill, showing that climatology-based approaches may be hard to beat due to the astronomical prevalence in determining solar production. In such cases, verifying the anomalies could reveal this earlier in the verification process.

4. From forecast skill to service value

In *ex ante* coevaluation, the alternative knowledge source is based on expert knowledge rather than a prior service that does not exist yet. This explains why alternative knowledge sources (Table 2c) and benchmarks (Table 2b) do not coincide in the case of SCHAT, SHYMAT, and SEAP and why skill (Fig. 2) and added value (Fig. 3) cannot be directly compared. This illustrates a challenge faced when trying to coordinate the verification and evaluation exercises at an early stage of service coproduction, but that can be addressed in *ex post* coevaluation.

Nevertheless, in the case of ROAT, a link appears between the skill of inflows and the final service value in terms of reservoir volumes (Fig. 3). ROAT forecasts have skill in sharpness, but lose reliability with respect to climatology (Fig. 2b) leading to a final service value which is worse than business-as-usual in 33% of the cases (wet years). Yet, ROAT reaches a 39% median gain toward the value of a service with perfect information particularly in dry years (cf. section 5). SHYMAT forecasts, however, gain in sharpness without sacrificing reliability (Fig. 2b) and their value is systematic and high (35% median gain). In the case of SCHAT, the service value in both Guavio and Betania (27% and 6% median gains, respectively) resembles their skill in accuracy (Fig. 2a). In Guavio, the forecast skill and service value are higher but more spread than in Betania, suggesting a possible link between skill and value despite discrepancies in benchmarks. A link between skill and value may thus be observed for services targeting the hydropower sector, potentially driven by more direct implications of forecast performance on economic gains in that sector (Cassagnole et al. 2021). Studies targeting reservoir operations and hydropower have further explored this relationship (Lee et al. 2022; Turner et al. 2017).

In supporting photovoltaic production, SEAP shows a systematic and high added value with respect to the actual collectors' tracking method (78% median gain), when its photovoltaic production forecasts show skill in correlation and no skill in accuracy compared to photovoltaic production simulated with averaged historical data (Fig. 2a). This may suggest that forecast correlation prevails for such applications.

5. Lessons learned through dialogues in the CLARA project

The dialogues between the service developers and users have highlighted several key messages. First, iterations in coevaluation played a role in shaping the services. For example, in SCHAT, the developers tested new algorithms to improve the inflow forecast skill, after the interactions

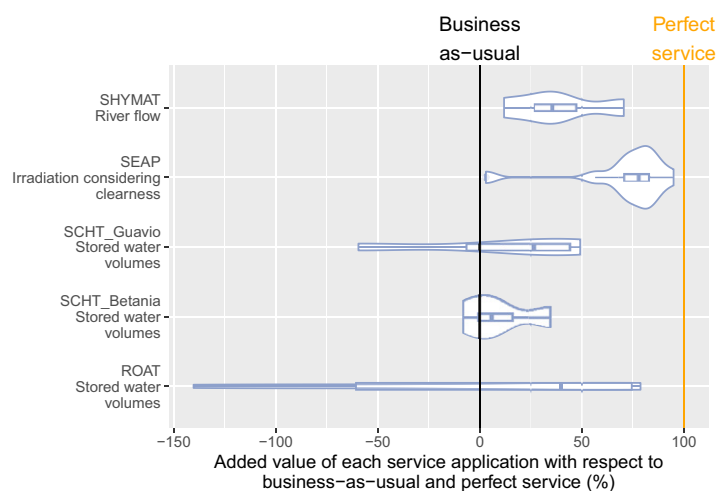


FIG. 3. Value of the services with respect to BaU value (black line) and to a service with perfect information (orange line). Violins (showing the density) and boxplots (showing quartiles) indicate the variability in value when the service supplies information for various years (all but SEAP) and months. WRI does not appear because its value could not be assessed due to limited in situ observations.

between socioeconomists and users, and a first round of evaluation. In the case of SHYMAT, these interactions led the developers to move from deterministic to probabilistic forecasts. A good representation of future uncertainties boosted service value in SHYMAT leading to higher informational benefits. Conversely, coevaluation may be an opportunity to demonstrate the role of uncertainties, especially since their inclusion in the decision chains remains a challenge.

Service value varies depending on the hydroclimatic conditions. Developers of ROAT explained the wide interannual variability in value by considering dry and wet years separately. In wet years, the service underestimated water volumes causing spillway discharge that highly penalized payoffs and value, while drier conditions were better predicted leading to high payoffs nearly equal to that of a perfect service. Skill and value exercises should thus demonstrate service potentials across hydroclimatic conditions to provide a comprehensive and nuanced picture. This retrospective exercise, however, requires long time series for robust analyses. Furthermore, service coevaluation can inform investment strategies in multisite optimization and management. The analysis of the SCHAT value suggests that investments to improve the service in the Betania reservoir would be more beneficial than in Guavio, since the maximum potential value is higher in the former location.

Limitations in the availability of in situ observations prevented a representative quantitative evaluation of the seasonal forecasts provided by WRI. This case thus illustrates a typical barrier to forecast verification and evaluation, which could be overcome over time by developing perennial measurement networks. Part of the data collecting process should thus ensure that the adequate observation networks are in place for such evaluations. In addition, short verification and evaluation periods lead to uncertain assessments, which should be interpreted with caution, could be quantified for instance through bootstrapping. Robust assessments thus require sustained efforts on data collection in time.

Last, quantitative evaluations only capture part of the final service value as the availability of the service alone may initiate changes in decision-making. The Land Reclamation and Irrigation Consortium of Burana has recognized benefits from WRI, with farmers observing their water use as the service information about potential saves became available. On the contrary, the potential value of ROAT was high, but due to large penalties when relying on forecasts that were not fulfilled, the effective service value was very low. This is in line with the current manager's way of acting and perception of seasonal information. Managers apply an undeclared safety factor in their decisions and act more conservatively than stated in the value exercise. For this reason, it will be necessary to increase the reliability of the seasonal forecasts for the managers to include them in their daily workflow.

6. Recommendations for joint verification–evaluation

Our experience within CLARA was that implementing a joint verification–evaluation exercise while ensuring interdisciplinarity and transdisciplinarity was challenging. The project time allowed the codevelopment of the services, forecast verification, ex ante coevaluation, but not ex post evaluation, and iterations were limited.

The verification and evaluation exercises took place over the 3 years of CLARA, but the interdisciplinary exchanges took place after. A comprehensive dialogue between disciplines (hydroclimate scientists and socioeconomists) and actors (data purveyors, service developers, and service users) should occur early on so as to coordinate the verification and evaluation and ensure a seamless service assessment. This can be facilitated through participatory approaches. Time and anticipation are essential to homogenize terminologies, clarify disciplinary objectives, and coordinate accordingly. In CLARA, multiuser forums involving developers and purveyors were organized every 6 months and allowed interactions, for instance, through serious games targeting verification or business models. Such initiatives should outlive research projects, at least at the service scale, to allow for long-term and iterative coevaluation.

Flexibility is required when screening methodologies to account for the needs and capacity of the users and developers. The coevaluation process within CLARA was fully adapted to the service decision-making contexts; verification, however, was partly designed to allow a cross-sector comparison between services. The comparison of both exercises revealed that the forecast variable being verified was not always the most relevant for users, the forecast benchmarks were sometimes unrealistic when forecasts were not already in place, and the forecast attributes considered for the evaluation were not always assessed in verification. A perspective would be to design a verification better aligned with the evaluation methodology and hence with user needs, and propose metrics focusing on the forecast attributes leading to value.

Last, in climate services, business models highlight key scientific advancements that increase value against competitors (Larosa and Mysiak 2020). The financial structure in particular (i.e., the costs and revenue architecture which were not considered here but analyzed within CLARA) can be used to secure the funding required for long-term and iterative coevaluation.

7. The proposed codevelopment and coevaluation framework

Based on this experience, we propose a framework that articulates well-known operational steps, namely, *screening*, *collecting*, and *evaluating* or *verifying* (Fig. 4). The actors in the framework are categorized as social and climate scientists, service users, and service developers, though these categories could overlap (e.g., climate scientists acting as service developers). The main recommendation is that actors in coevaluation and codevelopment should interact early on in the exercise.

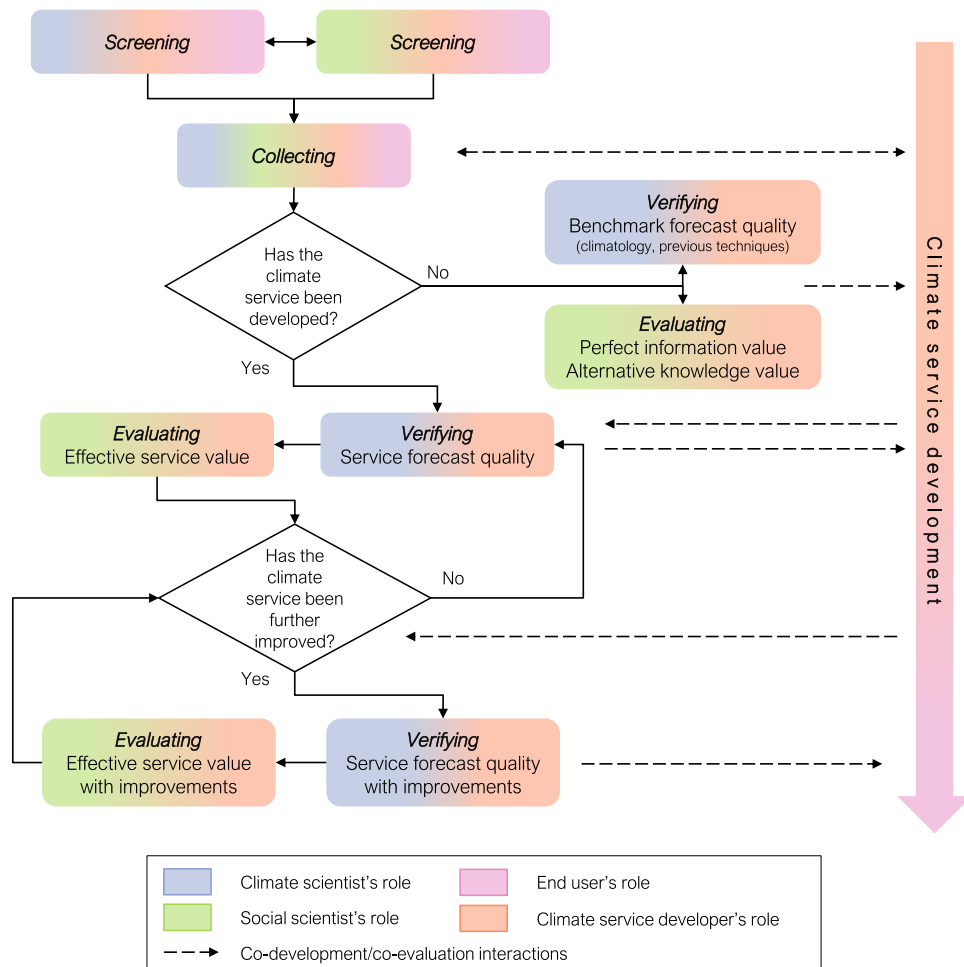


FIG. 4. Schematic representation of different phases of the forecast verification and service coevaluation processes.

Screening consists of reviewing the methodologies to verify forecast quality or assess the economic value of climate services. Social and climate scientists actively discuss pros and cons of each method with users and developers in light of the climate services' features to determine which one can capture most of their salient characteristics.

Collecting involves all actors, namely, climate and social scientists, climate services' producers, and users to gather information to apply the proposed methodology. This is a truly transdisciplinary phase (Lawrence et al. 2022): the social and climate scientists indicate the type of information required, the users populate the exercise with the information related to the case study (when evaluating or verifying) and decision-making process (when evaluating), and the climate service producers provide the forecasts and other technical information related to the service. This phase requires interactions between coevaluation and codevelopment, since users discuss aspects which may change their requirements.

Evaluating and verifying are iterative steps: effective forecast quality and service value are assessed once the service is developed and every time there is an update or improvement. Improvements may originate from the quality of the seasonal forecasts, or the dissemination and communication means. The updated service verification and evaluation thus inform on the improvements worth pursuing (e.g., advanced modeling techniques, data assimilation, and uncertainty communication).

Verification aims to assess forecast quality, which is process-, system-, location-, and time-dependent. The verification process is generally independent from the decision-making process, though it should ideally be designed to assess forecast attributes of relevance for the users. The quality of a forecast set is commonly assessed against a reference, e.g., observations from a long enough past period (or simulations in the absence of observations) to ensure robust quality estimates. It is then compared with a forecast benchmark, which either represents the best available forecast prior to the development of the service or a prior version of the service (Pappenberger et al. 2015; Wilks 2011).

When evaluating a climate service, value can be generated in different steps of the value chain, and thus may require different evaluation techniques (Perrels et al. 2013; Fig. 5). For service prototypes aiming to support operational decision planning, value originates not only from the quality of the provided forecasts but also from how well the information is tailored (for instance, the choice of the metrics presenting the economic service value) and communicated (Calvo et al. 2022) and how well the service supports decision-making and user needs (Materia et al. 2020). A quantitative evaluation as performed in CLARA could thus be complemented by a qualitative evaluation accounting for value drivers other than the forecasts.

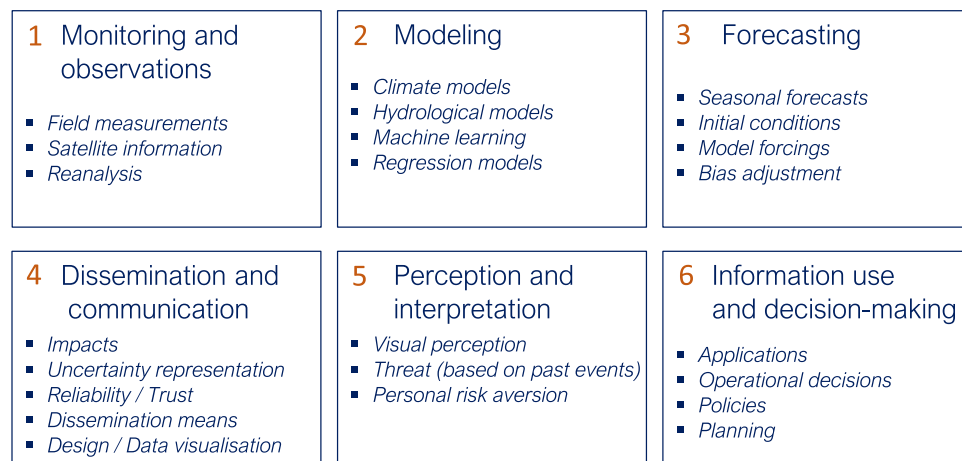


FIG. 5. A representative value chain for CSs, from monitoring and observations down to decision-making. Each step is identified and chained with upstream and downstream steps, while the examples of the methods or components involved in the investigated services are provided below each step.

8. Conclusions

The joint verification and evaluation exercise performed within CLARA across services highlights the diversity of situations that can be encountered when deploying the joint framework to support coproduction. The plurality in service designs reflects the plurality of decision-making contexts and user needs, which will also influence the skill–value relationship. Previous works (e.g., Cloke et al. 2017; Materia et al. 2020) suggest that improvements in early warning systems shall be assessed considering both skill and value. This work further suggests that (i) choices in the codevelopment of climate services should consider both forecast skill and service value; (ii) time and, therefore, sustained funding are key to allow for such interdisciplinary and transdisciplinary exercises; and (iii) verification exercises should evolve to target user needs in diverse sector-related decision-making contexts.

Acknowledgments. This research has been conducted with support from the European Union’s Horizon 2020 CLARA project under Grant Agreement 730482. L. Crochemore was funded by the Grenoble-INP/UGA starting grant and by Agence Nationale de la Recherche under Grant Agreement ANR-23-W4AP-0004-05. S. Materia has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement 101033654. I. G. Pechlivanidis was funded by the EU Horizon 2020 project Innovating climate services through integrating scientific and local knowledge (I-CISK) under Grant Agreement 101037293, and CLINT (Climate Intelligence: Extreme events detection, attribution and adaptation design using machine learning) under Grant Agreement 101003876.

Data availability statement. The seasonal forecasts used in the services are all freely available via Copernicus Climate Change Service. The services are meant for commercialization, and output data are thus sensitive, but some of the service platforms are openly accessible, such as SEAP, SHYMAT (paid access for access to some run-of-river systems), ROAT (paid access for access to some reservoir systems), WRI (login required), and SCHAT’s demo versions (restricted to nonsensitive data).

APPENDIX A

Formulation of Forecast Attributes and Skill

In the following equations, the notations are as follows: N is the number of years y , m is the forecast month, and lt is the forecast lead time. All metrics are then normalized so that their optimum value is 100.

a. Accuracy. The accuracy is measured with the mean absolute error, formulated as

$$MAE_{m,lt} = \frac{1}{N} \sum_{y=1}^N |F_{y,m,lt} - O_{y,m,lt}|, \quad (A1)$$

where $F_{y,m,lt}$ is the deterministic forecast or the mean (SCHAT and SEAP) or median (WRI) of the ensemble forecast and $O_{y,m,lt}$ is the reference observations for year y and month $m + lt - 1$.

b. Correlation. The correlation of forecasts with the reference is assessed based on the Spearman rank correlation coefficient. It is then multiplied by a factor of 100.

c. Reliability. Reliability is a characteristic of probabilistic forecasts that indicates the consistency between observed and forecast probabilities. In the literature, reliability is commonly based on a division of an ensemble forecast range into bins delimited by the ensemble members. Reliability can then be assessed based on how evenly the retrospective observation falls within each of these bins, for instance through a rank histogram (Anderson 1996).

Here, instead, the forecast is simplified to a single bin comprising the entire range between the 10th and 90th quantiles of the forecast ensemble for each time step (a single combination of year y , month m , and lead time lt). The reliability R , in this paper, is then formulated as the frequency with which the observation falls within this bin:

$$R_{m,lt} = \frac{100}{N} \sum_{y=1}^N \delta_{y,m,lt}, \quad (\text{A2})$$

where

$$\begin{cases} \delta_{y,m,lt} = 1, & \text{if } O_{y,m,lt} \in [Q_{10}(F_{y,m,lt}), Q_{90}(F_{y,m,lt})] \\ \delta_{y,m,lt} = 0, & \text{otherwise} \end{cases}$$

d. Sharpness. Sharpness indicates the spread of the ensemble forecast members with respect to the spread in historical observations (Wilks 2011). Here, it is expressed as the ratio between the 80% quantile range of the service-generated forecast (difference between the 90th and 10th percentiles of the forecast ensemble) against the 80% quantile range of climatology:

$$S_{m,lt} = \frac{\sum_{y=1}^N [Q_{90}(F_{y,m,lt}) - Q_{10}(F_{y,m,lt})]}{\sum_{y=1}^N [Q_{90}(C_{y,m,lt}) - Q_{10}(C_{y,m,lt})]}, \quad (\text{A3})$$

where $F_{y,m,lt}$ is the ensemble forecast and $C_{y,m,lt}$ is the ensemble based on historical observations for month m and lead time lt , excluding year y . It is independent from observations and, therefore, is not a quality indicator per se. Here, we consider it as complementary to the reliability indicator since “Sharp forecasts will be accurate only if they also exhibit good reliability” (Wilks 2011).

e. Accuracy in the case of binary events. The capacity of the ensemble forecasts to predict whether an event will occur or not (binary event) is evaluated based on the Brier score (Brier 1950). The occurrence of an event is defined based on the probability of reaching a fixed threshold. It is formulated as follows:

$$\text{Brier}_{m,lt} = \frac{1}{N} \sum_{y=1}^N (P_{F,y,m,lt} - P_{O,y,m,lt})^2, \quad (\text{A4})$$

where $P_{F,y,m,lt}$ is the forecast probability of the event occurring (between 0 and 1) and $P_{O,y,m,lt}$ is 1 if the event occurred and 0 otherwise.

f. Skill score. Skill scores compare any forecast characteristics or quality indicator to that of a reference forecast. Its most common formulation (Wilks 2011; Jolliffe and Stephenson 2003), ranging from 100 (optimum) to infinite negative values, is as follows:

$$\text{Skill}_{m,lt} = \frac{\text{Crit}_{m,lt}^{\text{CS}} - \text{Crit}_{m,lt}^{\text{BaU}}}{\text{Crit}_{m,lt}^{\text{PI}} - \text{Crit}_{m,lt}^{\text{BaU}}} \times 100\%, \quad (\text{A5})$$

where $\text{Crit}_{m,lt}^{\text{CS}}$ is the quality of the service-generated forecast, $\text{Crit}_{m,lt}^{\text{PI}}$ is the quality of a perfect forecast system, and $\text{Crit}_{m,lt}^{\text{BaU}}$ is the quality of the reference forecast. The skill score in forecast verification relates closely to the performance index in service evaluation (see appendix B).

APPENDIX B

Operationalization of the Value Assessment Methodology

Applying the value of information theory (Pope et al. 2017, 2019) requires inputs on the decision-making context through an indicator, a *decision space*, *states of the world*, *payoffs*, and forecast inputs of varying *skill*.

The indicator represents the link between the forecasts and the operative decisions. It could be the variable forecast by the service (as river flow in cubic meters per second for SHYMAT) or a variable derived from the forecast. For SCHAT, the service forecasts inflows to the reservoir, but the decision is taken based on the water volume in the reservoir. A reservoir management model was thus made available by the user to derive reservoir volumes from past inflows given reservoir characteristics. Similarly, SEAP produces forecasts of photovoltaic production, but states of the world are based on the daily irradiation on photovoltaic modules based on the clearness index (kt).

The decision space translates the user operative decisions. Since these services are mock-ups and decisions are not already in place, this decision space represents the users' decision to be supported once the service becomes operational. The decision space could consist of a "do–don't" decision, or be continuous as in the case of SHYMAT and SCHAT (setting the turbine level). Similarly, SEAP allows solar trackers to be programmed in advance; thus, the action is to position the solar trackers depending on the class of the clearness index. For ROAT, the decision on water management leads to three actions: "save water," "maintain water level in the dam," and "release water."

The user decision is expected to change depending on the state of the world the indicator falls into. These states can be constant in time (SCHAT and SHYMAT) or change monthly (ROAT). For SCHAT, the states of the world derive from the management rules of the dam and affect the level of turbine activation; for SHYMAT, instead, the states of the world derive from different levels of river flows, that ultimately affect the operation of the turbines. In the case of ROAT, some states are constant in time (as the flood control states); others change monthly based on the dam operating rules and the river basin management plan.

The payoffs associated with each action-state of the world combination can be presented in terms of a monetary or a physical unit of measure, or a different unit. In some cases, the value of the physical unit is unknown and the best way to proceed is to use a physical unit of measure acting as a proxy for savings or profits (as in SEAP). Because the presented services are prototypes and users were not always allowed to disclose potential profits, some users were asked to provide instead a score for combinations of action-forecasts-effective realizations. The final value was then given by the score. This methodology is applied for ROAT, SHYMAT, and SCHAT. ROAT's users define payoffs as linearly increasing inside the "normal state" of the world (the higher the water level, the higher the payoff), and as discrete in the other states of the world. Moreover, since the service works in a multipurpose context (multiobjective reservoir), payoffs are set by three different users before being weighed. SHYMAT and SCHAT users consider payoffs related to operational management only. However, they recognize that climate services could be used to assess, for instance, potential financial costs for the producer when entering the energy market or covering missing revenues by accessing credit markets (SCHAT) or to start machine repairs (SHYMAT).

Finally, the *performance* of the knowledge stocks represents the ability of a knowledge source to forecast states of the world. It relates to the quality assessment presented in this paper but considers which state of the world the hindcast predicts. The performance here consists of two contingency matrices, one for each knowledge stock. In this case, we are interested in the "right" forecasts (true negatives and true positives), in the "wrong" ones (false positives and false negatives), and in overestimations and underestimations, since each combination is linked to a potential course of action and associated payoffs.

The expected value is assessed for each knowledge stock, that is for the business-as-usual (BaU) knowledge and the climate service (CS) knowledge:

$$EV_{BaU} = \sum_{k=1}^N p_{BaU}(X_k) \min_A \sum_{n=1}^N \pi(A, x_n) P_{BaU}(x_n | X_k), \quad (B1)$$

$$EV_{CS} = \sum_{k=1}^N p_{CS}(X_k) \min_A \sum_{n=1}^N \pi(A, x_n) P_{CS}(x_n | X_k), \quad (B2)$$

where $p(X_k)$ is the forecast probability of event X_k according to the knowledge source (BaU or CS), $\pi(A, x_n)$ is the payoff associated with action A when the state of the world x_n occurs, and $P(x_n | X_k)$ is the probability of event x_n being observed when X_k is forecast.

The expected value of perfect information (EVPI) (Pope et al. 2017) is derived from the effective realizations in the test period:

$$EVPI = \sum_{n=1}^N \min_A \pi(A, x_n) p(X_n) - EV_{BaU}, \quad (B3)$$

also called “maximum potential value.” This is the upper bound for the climate service value.

Finally, we calculate the effective value of the climate service [or expected value of imperfect information (EVII)] as

$$EVII = EV_{CS} - EV_{BaU}. \quad (B4)$$

Moreover, following Richardson (2000), in each iterative phase of the evaluation when new improvements in the climate service are assessed, we calculate a performance index (PI) as

$$PI = EVII / EVPI. \quad (B5)$$

PI captures how much of the maximum potential value is effectively gained due to the development level of the climate service. PI is equivalent to the concept of forecast skill used in forecast verification (see appendix A).

Verification and evaluation terminologies

Terminologies are sometimes field-dependent. We thus define the terminologies used throughout the manuscript:

- *Forecast quality*: A measure of how correctly the forecast corresponds to the observations or simulations in the absence of observations (Murphy 1993).
- *Skill score*: A measure of the relative quality of a forecast set, assessed with respect to some reference forecasts, such as forecast climatology, persistence, and random forecast (Wilks 2011).
- *Service value*: Impact on user benefits of using climate information compared to using another knowledge source in decision-making (Meza et al. 2008).

Appendix A details the formulations of forecast attributes and skill, and appendix B details the formulations of service value.

References

- Alexander, M., and S. Dessai, 2019: What can climate services learn from the broader services literature? *Climatic Change*, **157**, 133–149, <https://doi.org/10.1007/s10584-019-02388-8>.
- Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1530, [https://doi.org/10.1175/1520-0442\(1996\)009<1518:AMFPAE>2.0.CO;2](https://doi.org/10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2).
- Baehr, J., and Coauthors, 2015: The prediction of surface temperature in the new seasonal prediction system based on the MPI-ESM coupled climate model. *Climate Dyn.*, **44**, 2723–2735, <https://doi.org/10.1007/s00382-014-2399-7>.
- Batté, L., L. Dorel, C. Ardilouze, and J.-F. Guérémy, 2019: Documentation of the METEO-FRANCE seasonal forecasting system 7. ECMWF, 30 pp., <https://www.umr-cnrm.fr/IMG/pdf/system7-technical.pdf>.
- Boon, E., S. J. Wright, R. Biesbroek, H. Goosen, and F. Ludwig, 2022: Successful climate services for adaptation: What we know, don't know and need to know. *Climate Serv.*, **27**, 100314, <https://doi.org/10.1016/j.cliser.2022.100314>.
- Brasseur, G. P., and L. Gallardo, 2016: Climate services: Lessons learned and future prospects. *Earth's Future*, **4**, 79–89, <https://doi.org/10.1002/2015EF000338>.
- Bremer, S., A. Wardekker, S. Dessai, S. Sobolowski, R. Slaattelid, and J. van der Sluijs, 2019: Toward a multi-faceted conception of co-production of climate services. *Climate Serv.*, **13**, 42–50, <https://doi.org/10.1016/j.cliser.2019.01.003>.
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78** (1), 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- Bruno Soares, M., M. Alexander, and S. Dessai, 2018: Sectoral use of climate information in Europe: A synoptic overview. *Climate Serv.*, **9**, 5–20, <https://doi.org/10.1016/j.cliser.2017.06.001>.
- Calvo, L., I. Christel, M. Terrado, F. Cucchiatti, and M. Pérez-Montoro, 2022: Users' cognitive load: A key aspect to successfully communicate visual climate information. *Bull. Amer. Meteor. Soc.*, **103**, E1–E16, <https://doi.org/10.1175/BAMS-D-20-0166.1>.
- Cantone, C., H. Ivars Grape, S. El Habash, and I. G. Pechlivanidis, 2023: A co-generation success story: Improving drinking water management through hydro-climate services. *Climate Serv.*, **31**, 100399, <https://doi.org/10.1016/j.cliser.2023.100399>.
- Cassagnole, M., M.-H. Ramos, I. Zalachori, G. Thirel, R. Garçon, J. Gailhard, and T. Ouillon, 2021: Impact of the quality of hydrological forecasts on the management and revenue of hydroelectric reservoirs—A conceptual approach. *Hydrol. Earth Syst. Sci.*, **25**, 1033–1052, <https://doi.org/10.5194/hess-25-1033-2021>.
- Clements, J., A. Ray, and G. Anderson, 2013: The value of climate services across economic and public sectors: A review of relevant literature. United States Agency for International Development, 26 pp., https://www.climate-links.org/sites/default/files/asset/document/CCRD-Climate-Services-Value-Report_FINAL_0.pdf.
- Cloke, H. L., F. Pappenberger, P. J. Smith, and F. Wetterhall, 2017: How do I know if I've improved my continental scale flood early warning system? *Environ. Res. Lett.*, **12**, 044006, <https://doi.org/10.1088/1748-9326/aa625a>.
- Contreras, E., J. Herrero, L. Crochemore, C. Aguilar, and M. J. Polo, 2020a: Seasonal climate forecast skill assessment for the management of water resources in a run of river hydropower system in the Poqueira River (southern Spain). *Water*, **12**, 2119, <https://doi.org/10.3390/w12082119>.
- , —, —, I. Pechlivanidis, C. Photiadou, C. Aguilar, and M. J. Polo, 2020b: Advances in the definition of needs and specifications for a climate service tool aimed at small hydropower plants' operation and management. *Energies*, **13**, 1827, <https://doi.org/10.3390/en13071827>.
- Crochemore, L., M.-H. Ramos, and I. G. Pechlivanidis, 2020: Can continental models convey useful seasonal hydrologic information at the catchment scale? *Water Resour. Res.*, **56**, e2019WR025700, <https://doi.org/10.1029/2019WR025700>.
- , C. Cantone, I. G. Pechlivanidis, and C. S. Photiadou, 2021: How does seasonal forecast performance influence decision-making? Insights from a serious game. *Bull. Amer. Meteor. Soc.*, **102** (9), 1–54, <https://doi.org/10.1175/BAMS-D-20-0169.1>.
- Delpiazzi, E., F. Bosello, P. Mazzoli, S. Bagli, V. Luzzi, and F. Dalla Valle, 2022: Co-evaluation of climate services. A case study for hydropower generation. *Climate Serv.*, **28**, 100335, <https://doi.org/10.1016/j.cliser.2022.100335>.
- Essenfelder, A. H., and Coauthors, 2020: Smart Climate Hydropower Tool: A machine-learning seasonal forecasting climate service to support cost–benefit analysis of reservoir management. *Atmosphere*, **11**, 1305, <https://doi.org/10.3390/atmos11121305>.
- Giuliani, M., L. Crochemore, I. Pechlivanidis, and A. Castelletti, 2020: From skill to value: Isolating the influence of end user behavior on seasonal forecast assessment. *Hydrol. Earth Syst. Sci.*, **24**, 5891–5902, <https://doi.org/10.5194/hess-24-5891-2020>.
- Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, <https://doi.org/10.1111/j.1467-9868.2007.00587.x>.
- Hargreaves, J. C., 2010: Skill and uncertainty in climate models. *Wiley Interdiscip. Rev.: Climate Change*, **1**, 556–564, <https://doi.org/10.1002/wcc.58>.
- Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, <https://doi.org/10.1002/qj.3803>.
- Johnson, S. J., and Coauthors, 2019: SEAS5: The new ECMWF seasonal forecast system. *Geosci. Model Dev.*, **12**, 1087–1117, <https://doi.org/10.5194/gmd-12-1087-2019>.
- Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons Ltd., 240 pp.
- Katz, R. W., and J. Lazo, 2011: Economic value of weather and climate forecasts. *The Oxford Handbook of Economic Forecasting*, Oxford Academic, 559–584, <https://doi.org/10.1093/oxfordhdb/9780195398649.013.0021>.
- , and A. H. Murphy, Eds., 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge University Press, 222 pp., <https://doi.org/10.1017/CBO9780511608278>.
- Larosa, F., and J. Mysiak, 2020: Business models for climate services: An analysis. *Climate Serv.*, **17**, 100111, <https://doi.org/10.1016/j.cliser.2019.100111>.
- Laugesen, R., M. Thyer, D. McInerney, and D. Kavetski, 2023: Flexible forecast value metric suitable for a wide range of decisions: Application using probabilistic subseasonal streamflow forecasts. *Hydrol. Earth Syst. Sci.*, **27**, 873–893, <https://doi.org/10.5194/hess-27-873-2023>.
- Lawrence, M. G., S. Williams, P. Nanz, and O. Renn, 2022: Characteristics, potentials, and challenges of transdisciplinary research. *One Earth*, **5**, 44–61, <https://doi.org/10.1016/j.oneear.2021.12.010>.
- Lee, D., J. Y. Ng, S. Galelli, and P. Block, 2022: Unfolding the relationship between seasonal forecast skill and value in hydropower production: A global analysis. *Hydrol. Earth Syst. Sci.*, **26**, 2431–2448, <https://doi.org/10.5194/hess-26-2431-2022>.
- Materia, S., Á. G. Muñoz, M. C. Álvarez-Castro, S. J. Mason, F. Vitart, and S. Gualdi, 2020: Multimodel subseasonal forecasts of spring cold spells: Potential value for the hazelnut agribusiness. *Wea. Forecasting*, **35**, 237–254, <https://doi.org/10.1175/WAF-D-19-0086.1>.
- Mauser, W., G. Klepper, M. Rice, B. S. Schmalzbauer, H. Hackmann, R. Leemans, and H. Moore, 2013: Transdisciplinary global change research: The co-creation of knowledge for sustainability. *Curr. Opin. Environ. Sustainability*, **5**, 420–431, <https://doi.org/10.1016/j.cosust.2013.07.001>.
- Meza, F. J., J. W. Hansen, and D. Osgood, 2008: Economic value of seasonal climate forecasts for agriculture. *J. Appl. Meteor. Climatol.*, **47**, 1269–1286, <https://doi.org/10.1175/2007JAMC1540.1>.
- Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, [https://doi.org/10.1175/1520-0434\(1993\)008<0281:WIAGFA>2.0.CO;2](https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2).
- Pappenberger, F., M. H. Ramos, H. L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, and P. Salamon, 2015: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *J. Hydrol.*, **522**, 697–713, <https://doi.org/10.1016/j.jhydrol.2015.01.024>.

- Pechlivanidis, I. G., L. Crochemore, J. Rosberg, and T. Bosshard, 2020: What are the key drivers controlling the quality of seasonal streamflow forecasts? *Water Resour. Res.*, **56**, e2019WR026987, <https://doi.org/10.1029/2019WR026987>.
- Perrels, A., Th. Frei, F. Espejo, L. Jamin, and A. Thomalla, 2013: Socio-economic benefits of weather and climate services in Europe. *Adv. Sci. Res.*, **10**, 65–70, <https://doi.org/10.5194/asr-10-65-2013>.
- Photiadou, C., and Coauthors, 2021: Designing a climate service for planning climate actions in vulnerable countries. *Atmosphere*, **12**, 121, <https://doi.org/10.3390/atmos12010121>.
- Pope, E. C. D., C. Buontempo, and T. Economou, 2017: Quantifying how user-interaction can modify the perception of the value of climate information: A Bayesian approach. *Climate Serv.*, **6**, 41–47, <https://doi.org/10.1016/j.cliser.2017.06.006>.
- , —, and —, 2019: Exploring constraints on the realised value of a forecast-based climate service. *Climate Serv.*, **15**, 100102, <https://doi.org/10.1016/j.cliser.2019.100102>.
- Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, <https://doi.org/10.1002/qj.49712656313>.
- Sanna, A., A. Borrelli, P. Athanasiadis, S. Materia, A. Storto, A. Navarra, S. Tibaldi, and S. Gualdi, 2016: CMCC-SPS3: The CMCC Seasonal Prediction System 3. CMCC, accessed 13 October 2021, <https://www.cmcc.it/publications/rp0285-cmcc-sps3-the-cmcc-seasonal-prediction-system-3>.
- Street, R. B., C. Buontempo, J. Mysiak, E. Karali, M. Pulquério, V. Murray, and R. Swart, 2019: How could climate services support disaster risk reduction in the 21st century. *Int. J. Disaster Risk Reduct.*, **34**, 28–33, <https://doi.org/10.1016/j.ijdr.2018.12.001>.
- Sutanto, S. J., F. Wetterhall, and H. A. J. Van Lanen, 2020: Hydrological drought forecasts outperform meteorological drought forecasts. *Environ. Res. Lett.*, **15**, 084010, <https://doi.org/10.1088/1748-9326/ab8b13>.
- Troccoli, A., 2018: Achieving valuable weather and climate services. *Weather & Climate Services for the Energy Industry*, A. Troccoli, Ed., Springer International Publishing, 13–25.
- Turner, S. W. D., J. C. Bennett, D. E. Robertson, and S. Galelli, 2017: Complex relationship between seasonal streamflow forecast skill and value in reservoir operations. *Hydrol. Earth Syst. Sci.*, **21**, 4841–4859, <https://doi.org/10.5194/hess-21-4841-2017>.
- Vaughan, C., and S. Dessai, 2014: Climate services for society: Origins, institutional arrangements, and design elements for an evaluation framework. *Wiley Interdiscip. Rev.: Climate Change*, **5**, 587–603, <https://doi.org/10.1002/wcc.290>.
- , M. F. Muth, and D. P. Brown, 2019: Evaluation of regional climate services: Learning from seasonal-scale examples across the Americas. *Climate Serv.*, **15**, 100104, <https://doi.org/10.1016/j.cliser.2019.100104>.
- Villani, G., F. Tomei, V. Pavan, A. Pirola, A. Spisni, and V. Marletto, 2021: The iCOLT climate service: Seasonal predictions of irrigation for Emilia-Romagna, Italy. *Meteor. Appl.*, **28**, e2007, <https://doi.org/10.1002/met.2007>.
- Weisheimer, A., and T. N. Palmer, 2014: On the reliability of seasonal climate forecasts. *J. Roy. Soc. Interface*, **11**, 20131162, <https://doi.org/10.1098/rsif.2013.1162>.
- Wilks, D. S., 2011: Forecast verification. *Statistical Methods in the Atmospheric Sciences*. Vol. 100, Academic Press, 301–394.