WATExR

Climate seasonal prediction for adapting water resources to extreme events

ERA4CS European Research Area for Climate Services

JPI Climate

WP 4 – Assessment

Quantification and visualization of combined input data and impact model skill at each site, and significance of this for decision-making

June 2021

Authors: François Clayer, Leah Jackson-Blake, Daniel Mercado-Bettín, Muhammed Shikhani, Tadhg Moore, Andrew French, James Sample, Maria-Dolores Frias, Sixto Herrera, Elvira de Eyto, Eleanor Jennings, Karsten Rinke, Rafael Marcé.

Table of Contents

1 Introduction

A key objective of the WATExR project is to integrate cutting-edge seasonal climate prediction and ecosystem impact models in forecasting tools tailored to the different needs of end-users in the water quality management sector. The water quality forecasting tools are designed to support decision making and adaptation strategies in water quality management. In WATExR deliverable 4.2, the forecasting skill of the tools was assessed for specific historic events selected together with the stakeholders, as well as for other combinations of season, impact variable and tercile for which impact models had significant skill (so-called “windows of opportunity”). Although very limited forecasting skill was reported for the historic events, a number of windows of opportunity were found. The source of the predictability for these windows of opportunity however remains to be investigated, and is the main focus of this deliverable.

A total of six forecasting setups for five case-studies are described and assessed below. These include catchment-lake process-based forecasts of inflow discharge and lake water temperature for four case-studies, one simple Bayesian network forecast for lake water quality and a statistical forecast of fish migration. In Part I of this deliverable, we describe the final setup for the forecasting tools at the various case-studies, present a set of experiments to investigate the origin of the forecasting skill, and finally discuss the results to better identify the source of the forecasting skills (Part I). This work is preliminary. Final results will be written up as a manuscript, to be submitted during 2021.

A seasonal forecast has no value without an indication of how much the predictions can be trusted. Clear communication of uncertainty is thus required to adequately use seasonal climate services for water management. In order to make sure that tool developers and stakeholders understand the different uncertainties in the forecasts, we carried out a workshop on how to best communicate forecasting uncertainty in the tool layouts. Activities and outcomes of this developer-stakeholder interaction workshop are presented in Part II.

Part I: Sources of forecasting skill

1. Description of the forecasting tool setup

The case study sites and modelling chain setups have already been described in Deliverable 2.3 and in Mercado-Bettin et al. (2021), although Deliverable 2.3 does not include final adjustment of the workflows, and so a final overview is presented below.

1.1 Case study sites

Characteristics of the five case study sites included in this report, including three water reservoirs, one lake and one catchment, are given in Table 1.

Table 1: Lake characteristics at the study sites, forecasting approach and forecasted variables.

Case study (Country)	Surface area (ha)	Volume (hm^3)	Water retention time (yr)	Max. Depth (m)	Mixing regime	Forecasting approach	Forecasted variables
Sau (Spain)	575	165	0.2	60	mono.	Catchment-lake process-based model	Discharge Surface & bottom Temperature
Mt Bold (Australia)	254	46.4	0.2-0.6	44.5	mono.	Catchment-lake process-based model	Discharge Surface & bottom Temperature
Vansjø (Norway)	3600	252	1.1	19	di.	Catchment-lake process-based model	Discharge Surface & bottom Temperature
						Simple Bayesian network/Naïve model	TP Chl-a Cyano Colour
Wupper (Germany)	211	26	0.2	31	di.	Catchment-lake process-based model	Discharge Surface & bottom Temperature
Burrishoole (Ireland)	Catchment described in the text					Catchment process-based and correlative statistical model chain	Migration timing of juvenile salmonids and maturing eels

1) Sau Reservoir (Spain) is a recreational and water supply system for the Barcelona metropolitan area of the Ter River catchment (1680 km^2) with a mean inflow of 14 $m^3 s^{-1}$.
2) Mount Bold (Mt. Bold) reservoir is the largest reservoir in South Australia. It receives water from the Onkaparinga catchment (325 km^2) and the Echunga Creek Catchment (32 km^2). In addition to both these inflows, the Onkaparinga river is supplemented with water from the Murray River via a pipeline providing water during the summer and autumn seasons where there is little to no precipitation. Mt. Bold reservoir supplies water to a drinking water reservoir for Adelaide and the surrounding Mount Lofty Ranges.
3) Lake Vansjø (Norway) is a drinking water source for three municipalities (about 60000 inhabitants) and is a major recreational and fishing area in the region. The lake is composed of several subbasins of which the two largest are Storefjorden (eastern basin,

sub-catchment of 244 km^2, surface area of 23.8 km^2), and Vanemfjorden (western basin, sub-catchment of 58 km^2, surface area: 12.0 km^2). The water flows through the deeper Storefjorden basin (max depth: 41 m, mean depth: 8.7 m, and residence time: 0.85 year) through a shallow channel to the shallower Vanemfjorden basin (max depth: 19.0 m, mean depth 3.8 m, residence time 0.21 year).

4) The Wupper reservoir is located in the West of Germany near Cologne and is supplied by the Wupper river with a catchment of 215 km^2.
5) The Burrishoole catchment (area ~84 km^2), west Ireland, is dominated by blanket peat and mountains (~ 600 m max. elevation) delineate its area, which ultimately drains into the largest freshwater lake in the catchment, Lough Feeagh (area: ~4 km^2). At least 70 km of small (< 10 m width) river channels drain into Lough Feeagh. Smaller lakes, the largest of which is Bunaveela (area: 1.4 km^2), are found in the upper catchment and form part of a navigable stream network (total area of ~0.22 km^2) accessible to diadromous fish. From Lough Feeagh, two channels (< 200 m) drop 10 m into Lough Furnace, a coastal lagoon with a tidal connection to the Atlantic Ocean. These two channels, the Salmon Leap and the Mill Race, contain upstream and downstream fish traps, which have captured largely complete daily censuses of migrating salmon and trout smolts (spring migration) and silver eels (autumn migration) leaving freshwater since 1970.

1.2 Climate data

Briefly, we used two different climate datasets to force the impact models in our workflows, a climate reanalysis and a seasonal climate forecasting product which are described in detail by Mercado-Bettín et al. (2021):

1) ERA5 is the latest reanalysis (Hersbach et al., 2020) produced by the European Centre for Medium-Range Weather Forecasts (ECMWF, https://www.ecmwf.int) within the Copernicus Climate Change Service (C3S, https://climate.copernicus.eu/). ERA5 data (1988-2016) was used to provide climate pseudo-observations for retrospective skill evaluation of seasonal climate forecasts and to correct for bias in the seasonal forecasting data.
2) SEAS5 is the latest seasonal forecasting system (Johnson et al., 2019) provided by the ECMWF. This forecasting system provides (i) real-time seasonal forecasts and (ii) retrospective seasonal forecasts for past years (hindcasts). We only used retrospective seasonal forecasts (hindcasts; 1994-2016) to validate our workflows. A hindcast with 25 members was considered for the period 1993-2016. The seasonal forecast covers up to the next 7 months, including present month. We used the SEAS5 data for the boreal seasons (spring: March through May; summer: June through August; autumn: September through November; winter: December through February), with one month as lead time (i.e. forecasts are initialized one month in advance of the target season: in November for winter, in February for spring, in May for summer and in August for autumn). SEAS5 data were bias-corrected as described in Mercado-Bettín et el. (2021).

Both climate datasets include air temperature (tas), wind speed (u and v components; uas and vas), air pressure (psl), relative humidity (tdps), cloud cover (tcc), solar radiation (rsds), and precipitation(tp). The seasons were defined as shown in Table 2 below, except for simple Bayesian network/naïve modelling of water chemistry at Vansjø (see Section 1.5) where only one season was defined as a composite of early and late summer, i.e., May to October, consistent with water quality classification scheme used in the Water Framework Directive (WFD) implementation in Norway. Note that for the WFD growing season, i.e., May to October, none of the SEAS5 climate forecast showed any significant skill.

An evaluation of the skill of the SEAS5 forecast for the different case study regions is presented in Mercado-Bettín et al. (2021) as well as in Deliverable 4.2, which shows that the forecasts have limited skill in predicting most of the variables of interest in the study regions. These assessments were carried out with the Relative Operating Characteristic Skill Score (ROCSS) which is a common measure to evaluate the skill of the probabilistic forecast (Jolliffe and Stephenson, 2003; Mason, 2013). The relative operating characteristic (ROC) is based on the ratio between the hit rate and the false alarm rate and is evaluated at each tercile separately. ROCSS evaluates the relative improvement of ROC by the forecast with respect to long-term climatology, and ranges from 1 (perfect discrimination) to -1 (perfectly bad discrimination), a zero value indicating no skill compared to a long-term climatology. The significance level against the null hypothesis (no skill) was calculated using Monte-Carlo techniques.

Table 2: Skills of the SEAS5 climate forecasts at each case-study as assessed by ROCSS against pseudo-observations. From Mercado-Bettín et al. (2021).

Site	Numbers of skilful SEAS5 climate forecasts variable - tercile				
	Winter	**Spring**	**Summer**	**Autumn***	**Total**
Norway	3 tcc - below rsds - above rlds - normal	7 psl – normal psl – above tas – above tcc - normal tdps – below uas – above vas - below	0	0	10/96
Australia	1 cc - normal	5 psl - above tcc - above tas - normal rsds - above petH - above	2 psl – above tdps - above	1 rlds - above	9/96
Spain	0	2 tcc – above psl - above	2 tcc – above tdps - above	1 tcc - above	5/96
Germany	2 rlds – normal vas - below	1 tdps - above	0	0	3/96
Ireland	-	0	-	0	0/18

1.3 Observations

Model forecasting skill was assessed by comparing modelled outputs to observations. The following observed time series were available for comparison at each site:

Table 3: Time series of observations available at each case-study.

Site	Variable	Start	End	Big gaps	Number of seasons with data	Sampling season/frequency
Norway	TP Chl-a Colour Cyanobacteria Discharge Temperature	1980 1980 1982 1996 1993 1996	2020 2020 2020 2020 2018 2016	1999 1999 1998-9, 2001-04 None 2001-04	37/38 37/38 36/38 34/38 100/100 54/100	Apr-Oct (weekly-monthly) Apr-Oct (weekly-monthly) Apr-Oct (weekly-monthly) Apr-Oct (weekly-monthly) All year (daily) Apr-Oct (weekly-monthly)
Spain	Discharge Temperature	1997	2018	2006		daily weekly-monthly
Australia	Discharge Temperature	2003 2006	2013 2015	2009-2013		daily daily-weekly
Ireland	Fish counts Discharge Water temperature	1979	2019	1984 and 1989	39 years (37 for eel)	daily
Germany	Discharge Temperature chemistry biology	1990	2018	none	28	daily once or twice every month once or twice every month once or twice every month

1.4 Catchment-lake process-based modelling

1.4.1 Models used

A catchment-lake process-based model chain was set up at four of the case-studies (Spain, Australia, Norway and Germany) to predict inflow discharge into the lake/reservoir and lake water temperature (Figure 1).

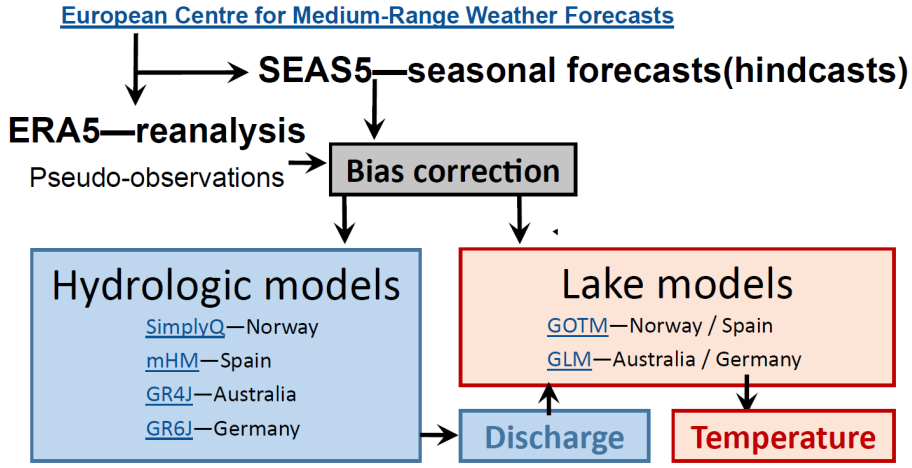

Figure 1: Workflow to produce discharge and lake water temperature forecasts. Calibrated hydrologic and lake models are used to produce seasonal forecasts with 25 climate seasonal members

Four hydrologic models were used, one for each case study. The mesoscale Hydrologic Model (mHM v5.9: http://www.ufz.de/mhm) was used to implement the hydrologic simulations in the Ter River catchment, which is the main inflow for Sau Reservoir. The *Génie Rural* (GR) suite of models implemented within the R package airGR (Coron et al., 2017) were used to model the inflows for the Wupper Reservoir and the Mt. Bold Reservoir (Onkaparinga and Echunga

Creek), namely GR6J and GR4J, respectively. The hydrologic module of the SimplyP catchment model for phosphorus, SimplyQ, was used to model the inflows to Lake Vansjø (Norway), and is described in detail by (Jackson-Blake et al., 2017). All hydrologic models run at a daily time step and were calibrated and validated using the Nash–Sutcliffe efficiency coefficient (NSE) as objective function. Hydrologic models were forced with precipitation (tp) and mean (also minimum and maximum in some cases) daily air temperature (tas).

The General Ocean Turbulence Model (GOTM, http://gotm.net) was used for simulating the thermal dynamics of Sau Reservoir (Spain) and Lake Vansjø (Norway). The General Lake Model (GLM, Hipsey et al., 2019) was used for simulating the thermal dynamics of the Mt. Bold and Wupper reservoirs. Lake models were calibrated and validated using the root mean square error (RMSE) and/or NSE as objective function(s). Lake models were forced with air temperature (tas), wind speed (u and v components; uas and vas), air pressure (psl), relative humidity (tdps), cloud cover (tcc), solar radiation (rsds), and precipitation(tp).

Regarding the lake water balance, we opted for a different parametrization of the models applied to reservoirs (Australia, Spain and Germany) versus the lake (Norway). For Lake Vansjø, the water level was simply fixed, which is usually a reasonable assumption since water level fluctuations are restricted to less than 1 m. Such a water level variation is not critical for the heat budget of the lake. Note also that, even though Lake Vansjø is a drinking water source, the amount of water extracted for use is negligible. For the reservoirs, on the other hand, much larger water level fluctuations are observed because complex and large variations in water pumping patterns occur due to the relative scarcity of water resources. It was thus critical to allow for water level fluctuations and parametrize the outflows to avoid dry-outs and to reflect water authorities' operations. The most realistic approach was to use a simple linear regression between observed inflow and outflow to be able to predict the outflow in the absence of observations.

1.4.2 Calibration and validation

Both hydrologic and lake models were calibrated against local observations, using ERA5 reanalysis data as forcing meteorological data. The resulting hydrologic and lake/reservoir simulations were highly consistent with real observations. Clear independent calibration and validation periods were defined for each case-study. Most common statistical goodness-of-fit parameters, e.g., Kling-Gupta efficiency (KGE), NSE and RMSE, for hydrological and lake modelling are reported in Tables 4 and 5, respectively.

Table 4: Goodness of fit statistics for the hydrologic model for each case-study

Country	River	Model	Warm-up	Calibration			Validation		
				Time	NSE	KGE	Time	NSE	KGE
Spain	Ter	mHM	5 years	1997-2007	0.60	0.66	2008-2018	0.54	0.63
Australia	Echunga	GR4J	5 years	2003-2007	0.64	0.70	2008-2013	0.80	0.75
	Onkaparinga	GR4J	5 years	1999-2002	0.80	0.84	2003-2006	0.65	0.54
Norway	Vansjø	SimplyQ	5 years	2005-2010	0.51	0.56	2011-2015	0.57	0.57
Germany	Wupper	GR6J	1 year	1991-2011	0.71	0.85	2012-2016	0.63	0.81

* Calculated from daily discharge

Table 5: Goodness of fit statistics for the lake model for each case-study

Country	Lake	Model	Warm-up	Calibration			Validation		
				Time	NSE	RMSE	Time	NSE	RMSE
Spain	Sau	GOTM	1 year	1997-2007	0.93	1.63	2008-2018	0.94	1.45
Australia	Mt. Bold	GLM	1 year	2014-2016	0.91	1.17	2016-2018	0.78	1.50
Norway	Vansjø	GOTM	1 year	2005-2010	0.92	1.12	2011-2015	0.93	1.10
Germany	Wupper	GLM	1 year	1993-2010	0.93	1.31	2011-2016	0.91	1.53

* Calculated from daily surface water temperature

Although we used a variety of hydrologic and lake models, common methods and codes were established to manipulate input and output data to and from the various models.

1.4.3 Generating forecasts for the hindcast period

For every modelled season during the hindcast period (time range 11/1993-11/2016, i.e., 92 runs from 23 years x 4 seasons), we implemented the following procedure to simulate seasonal river discharge and surface and bottom water temperature ensemble predictions (illustrated for Lake Vansjø in Figure 2):

1. Impact models (hydrologic and lake) were warmed up using ERA5,
2. A 4-month long simulation was run driven by the 25 ensemble members from the bias-corrected SEAS5 hindcast set for each initialization considered (e.g., February for spring).
3. Hydrologic and lake model outputs for the final 3 months (March to May) corresponding to the target season were selected for calculation of probabilistic expectations, while the initialization month was removed from the analysis since it is considered as a transitory period.

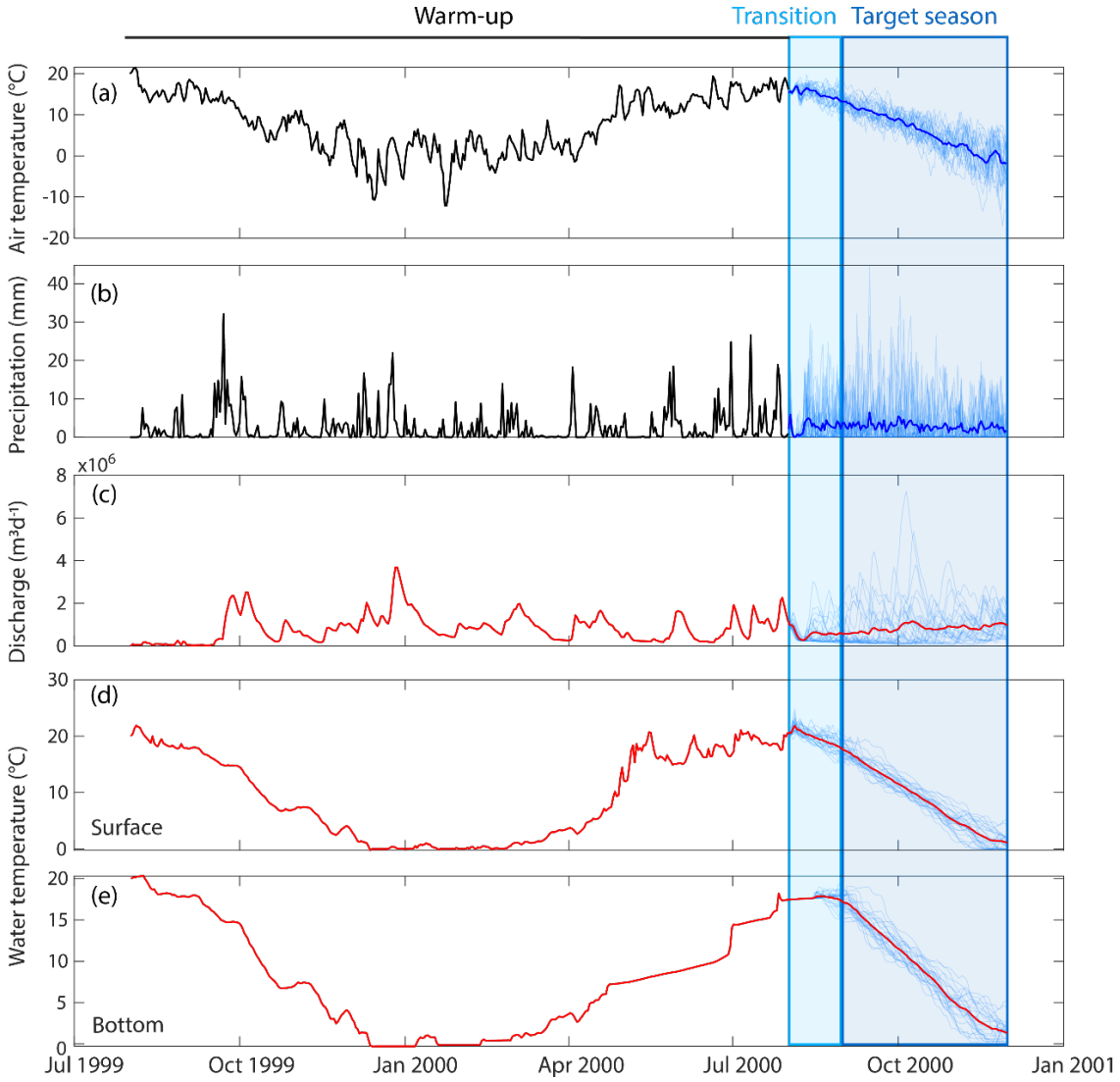

Figure 2: Time series of air temperature (a) and precipitation (b) from ERA5 data (black line) followed by bias-corrected SEAS5 forecasts (mean in dark blue, light blue lines are the 25 members) for Autumn 2000 (including the transition month), as well as modelling outputs of discharge (c) and lake surface (d) and bottom (e) temperature. Only one year is shown for the warm-up period for better readability of the rest of the periods.

1.5 Simple Bayesian network/Naïve modelling (Norway)

In Lake Vansjø, the water manager was also interested in forecasts of lake ecological status (according to the WFD classification scheme). A number of empirical models were therefore developed to predict WFD-relevant water quality/ecology parameters of interest. Four variables are predicted: mean concentrations of total phosphorus (TP), chlorophyll-a (chl-a) and lake colour, and maximum values of cyanobacteria biovolume (consistent with WFD classification criteria). Mean colour is forecasted as it may affect cyanobacteria and also be of interest in terms of drinking water treatment. Forecasts are produced for the period May-October, to coincide with the definition of the growing season used in WFD classification.

A number of empirical models were explored, starting from the simplest possible benchmark model, and increasing in complexity as more predictor variables were added:

Model 1*: Seasonal naïve forecast*: the simplest benchmark model. Each forecast is simply the value observed during the previous summer.
Model 2: *Gaussian Bayesian Belief Network, no weather nodes*: This is the same as model 3 (BN; Figure 3), except meteorological conditions are set to the long-term average and kept constant when making predictions. This is equivalent to removing weather nodes from the BN structure shown in Figure 3.
Model 3: *Gaussian Bayesian Belief Network, including weather nodes:* observed (and for future projections, forecasted) meteorological data are included when making predictions.

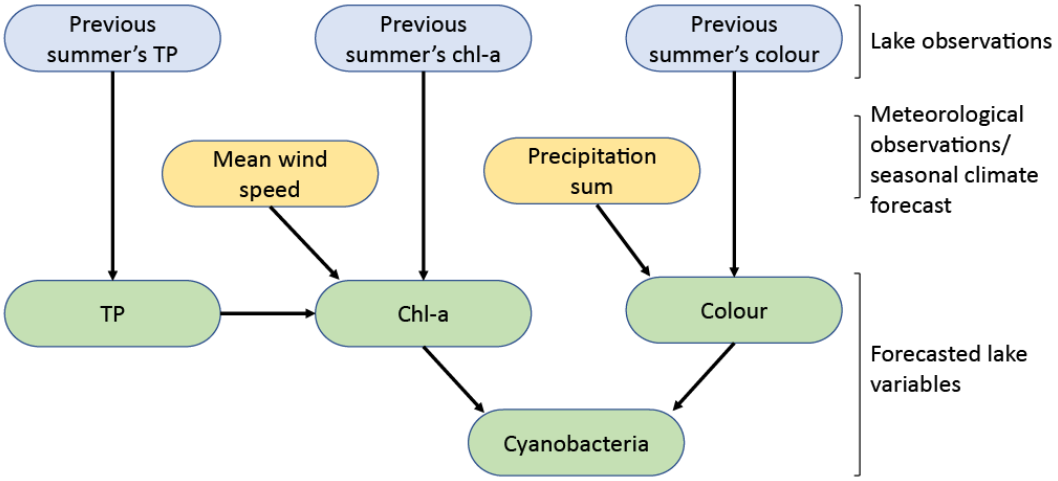

Figure 3: Bayesian Network (BN) structure considered in Model 2 (where yellow meteorological nodes were excluded) and Model 3. Green nodes are the end-points of interest.

Historic skill was assessed by comparing predictions and observations over the period 1981-2019 for TP, colour and chl-a, and 1996-2019 for cyanobacteria (for which fewer data were available). Skill was assessed using leave-one-out cross validation and a combination of skill metrics (predictive correlation, RMSE, classification error, Matthew's correlation coefficient (MCC)). The model with the highest predictive skill was then chosen for operational forecasting, as follows:

- TP, cyanobacteria, colour: Model 2, i.e. Gaussian BN, no weather nodes.
- chl-a: Model 1, seasonal naïve forecast

Skill scores for the models chosen for each variable, derived from leave-one-out cross validation, are given in Table 6. Qualitative thresholds of MCC>0.5 and classification error<0.2 (i.e. forecasts were wrong less than 20% of the time during the historic period) were chosen to select models whose performance was considered sufficient to inform management. The result was forecasts for chl-a and cyanobacteria.

Table 6: Forecast skill statistics for lake water quality forecasting model(s) used in Vansjø

Variable	Model	Pearson's R	RMSE	MCC	ROCSS	Classification error
Chl-a	Seasonal naïve	0.65	4.6	0.71	0.70	0.11
Colour	BN	0.82	9.4	0.47	0.46	0.23
Cyano	BN	0.67	0.95	0.78	0.76	0.13
TP	BN	0.58	3.9	0.34	0.34	0.33

1.6 Forecasting the timing of fish migration (Ireland)

In Ireland, an empirical model was developed to forecast the timing of seaward migration of diadromous fish. Explanatory variables in this model included simulated discharge, generated by a process-based hydrology model (GR4J), as well as a number of other variables that influence the preparedness of fish for migration. Details of the workflow have not been presented in previous deliverables, and are provided in Appendix 1.

1.6.1 Fish count model validation under known conditions

For each species, bias-corrected ERA5 was used as forcing data for the model chain to gauge predictive performance under known conditions. Here, we substituted bias corrected ERA5 in place of each of the 25 members of SEAS5 for each year and species and generated simulations for each species. This approach produced tercile based anomaly expectations for migration summary statistics under known environmental conditions (e.g., probabilities that the median day of migration is early, late or average; see Fig 4), which facilitated comparison with observed anomalies in a way that is consistent with probabilistic seasonal forecast inference.

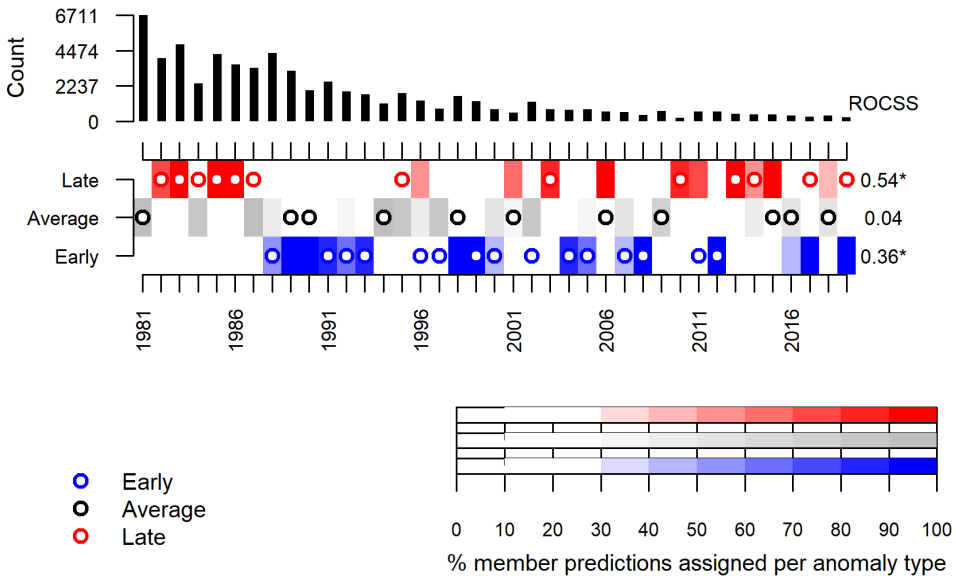

Figure 4: Predictive performance of trout correlative model under known weather. We simulated counts from 25 identical ERA members to produce a probabilistic expectation for each anomaly and year where weather conditions are known.

1.6.2 Fish count model validation under operational (re-)forecast scenarios and identifying windows of opportunity

For each year and correlative model predictor variable, we appended bias corrected lead-in ERA5 data to each of the 25 bias corrected SEAS5 seasonal forecast members. That is, for salmon and trout smolts, we appended ERA5 from the winter solstice (December 21/22) up to and including January 31 to SEAS5 for February 01 to August 31; whereas for silver eels, we appended ERA5 from the winter solstice up to and including June 30 to SEAS5 for July 01 to December 31. The resultant appended 25 datasets per year constituted multi-member forcing data to generate fish count simulations (see Fig 5) from which we calculated migration timing summary statistics and probabilistic forecast anomaly expectations (see Fig 6).

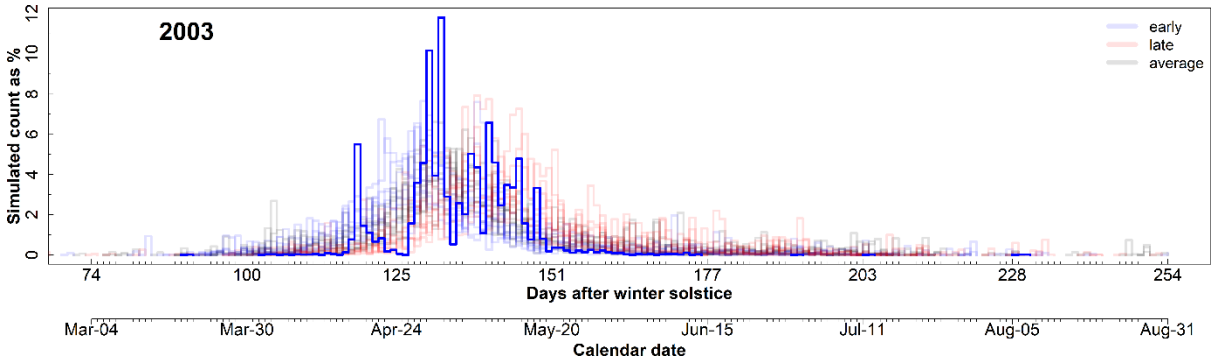

Figure 5: Simulated counts generated by correlative salmon model forced by 25 members of SEAS5 subsequently used to calculate probabilistic expectations. Migration timing summary statistic used to stratify simulations by tercile anomalies is the mean date of migration, which was calculated from all counts in each simulation that occurred between March and June. Bold blue line represents the observed counts for the early migration of 2003.

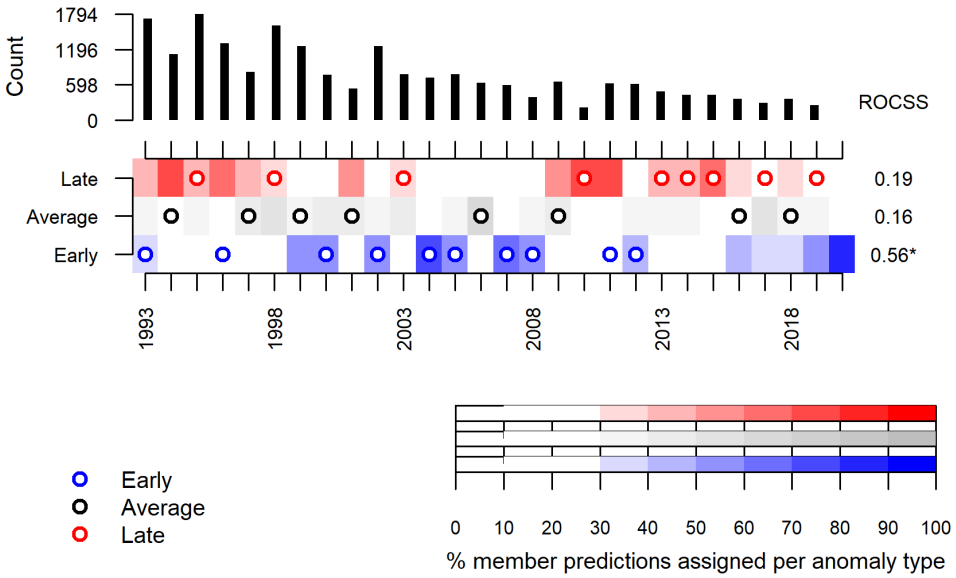

Figure 6: Evaluation of predictive performance of model chain for trout under operational (re-) forecast scenarios.

Owing to the probabilistic nature of seasonal forecast inference, operational skill for forecasting a selected anomaly should not be considered a window of opportunity for forecasting fish migration timing unless similar or greater skill also arises under an ideal

forecast scenario. In this case, we only identify one such window, wherein skill for forecasting the early anomalies for median year day of trout migration is maintained between ideal and operational scenarios (see comparative terciles in Figs 4 and 6). Moreover, such a skilful forecast should gain stronger credibility if probabilistic expectations of individual year-by-year comparisons between ideal and operational scenarios for that tercile are consistent.

2 Methods for assessing forecast uncertainty and sources of skills

2.1 Model outputs produced and modelling error

Table 7 displays the basic model outputs used in the assessment. These model outputs were then compared to various data, as shown in Table 8, to evaluate the difference sources of uncertainty in the model chains. Statistics used includes correlation coefficients, NSE, KGE and bias, as well as tercile plots and ROCSS when a probabilistic ensemble was generated (e.g., when SEAS5 data were used as forcing data).

Table 7: Original model outputs produced

Name	Model (Section)	Forcing meteorological data	Purpose
ERA5_lake	Catchment-lake process-based model (Section 1.4)	ERA5	Forecast that would be produced assuming a perfect lake model
SEAS5_lake	Catchment-lake process-based model (Section 1.4)	SEAS5	Full seasonal climate forecast, representative of what would be produced in any operational tool
ERA5_fish	Catchment process-based and correlative statistical model chain (Section 1.6)	ERA5	Probabilistic predictions from complete model chain that would be produced under a “perfect” forecast scenario.
SEAS5_fish	Catchment process-based and correlative statistical model chain (Section 1.6)	SEAS5	Full seasonal climate forecast, representative of what would be produced in any operational tool
SiNaï	Simple Bayesian network/Naïve modelling (Section 1.5)	none	Operational forecast

Table 8: Comparison performed to evaluate model uncertainty. See Table 7 for a description of the outputs used.

Expt	Outputs used	Data to evaluate forecast performance	Purpose
1	SEAS5_lake SEAS5_fish	ERA5_lake ERA5_fish	Transfer of climate model forecast skill through lake/fish model (assumes perfect impact models)
2	ERA5_lake ERA5_fish	Lake observations Fish observations	Predictive skill of model chain under perfect forecast
3	SEAS5_lake SEAS5_fish	Lake/Fish observations	Total forecasting skill of impact models chain (including error from seasonal forecasts and impact models)
4	SiNaï	Lake observations	Skill of BN/naïve forecasting system

2.2 Investigation of the source of forecasting skills

Within deliverable 4.2, we carried out a comprehensive assessment of which seasons and variables the forecasting systems have skill in (what we term “windows of opportunity”). Since then, a number of improvements have been made to several of the workflows, and so an updated set of windows of opportunity are displayed in Table 9. Note that the 5% significance level used means that some of may be false positives.

To identify the origin of the forecasting skill identified within the windows of opportunity in Table 9, we performed a set of experiments described below. Two sets of experiments were used which we thereafter qualified as “probabilistic” and “deterministic”, described as experiment series A and B in Table 10 below, respectively. Both sets of experiments were aimed at quantifying how much of the forecasting skills originates from the **boundary conditions** (forcing SEAS5 data during the target season; row 2 in Table 10) or from the **initial conditions** (forcing ERA5 data during the warm-up and eventually forcing SEAS5 data during the transition month; rows 3 and 4 in Table 10). In addition, a probabilistic experiment was performed to quantify the sensitivity of the modelling output to each specific variable in the boundary conditions (one specific variable in the forcing SEAS5 data during the target season; row 1, experiment 1A in Table 10).

Table 9: Catchment/lake variables, seasons and classes for which seasonal surface water forecasts had skill, as assessed by comparison to surface water (pseudo-)observations from the catchment/lake.

Site	Number of skilful /total combinations[a]	Catchment/ lake variable	Season	Class	Statistic(s) used in assessment of historic skill	Historic assessment period
Norway	9/36	Discharge	Spring	Lower, Upper	ROCSS	1993-2016
		Surface Temperature	Winter	Lower		
			Spring	Lower, Upper		
		Bottom Temperature	Winter	Lower, Upper		
			Spring	Lower, Upper		
	2/4	chl-a	Growing season	Lower, Upper	MCC, ROCSS, classification error	1981-2019
		cyanobacteria				
Australia	1/36	Echunga Discharge	Spring	Lower	ROCSS	1993-2016
		Bottom temperature	Winter	Lower		
			Summer	Lower		
Spain	10/36	Discharge	Summer	Lower	ROCSS	1988-2020
			Autumn	Upper		
			Winter	Normal		
		Surface Temperature	Summer	Lower, Upper		
		Bottom Temperature	Spring	Upper		
			Summer	Lower, Upper		
			Autumn	Lower, Upper		
Germany	1/36	Bottom Temperature	Spring	Lower, Upper	ROCSS	1993-2016
			Summer	Lower, Upper		
Ireland	3/72	Salmon	Spring migration	Day when 5% migrated: late	ROCSS	1993 - 2019
		Trout		Median day of migration: early		
		Eel		Day when 25% migrated: late		

[a] For all case studies except water quality variables in Norway, ROCSS scores were calculated for 36 data 'slices' in total: 3 impact variables x 4 seasons x 3 terciles. For water quality variables in Vansjø, scores for 4 impact variables for only one season were calculated.

The **probabilistic** experiments (A in Table 10) consisted in replacing the forcing data of interest, i.e., either boundary or initial conditions, by forcing data from an equivalent season but from a randomly selected year.

Experiment 2A: In the case of the probabilistic experiment for boundary conditions, for example Autumn 2000, the SEAS5 forcing data covering the target season, i.e., September to November 2000, is replaced by SEAS5 forcing data from a randomly selected Autumn (September to November, e.g., Autumn 2005). To ensure that the randomly sampled SEAS5 data are representative of the whole SEAS5 dataset, we introduce two levels of repetitions. First, when forecasting a specific season with random SEAS5 data, we randomly selected a year for each of the 25 members of SEAS5, meaning that the selected data to replace the

original SEAS5 forcing data is extremely likely to be from a different year for each SEAS5 member. Second, we repeated the experiment at least five times, and up to 25 times, for each season.

Experiment 3A: In the case of the probabilistic experiment for the initial conditions, for example Fall 2000, the ERA5 forcing data covering the warm-up period, i.e., August 1999 to July 2000, is replaced by randomly-selected ERA5 data.

Experiment 4A: In this experiment, the transition month as well as the warm-up period were replaced by randomly-selected ERA5 data. E.g. for an autumn (Sep-Nov) 2000 forecast, August 1999 to August 2000 were replaced by random ERA5 data. We also repeated this experiment at least five times, and up to 25 times, for each season.

The **deterministic** experiments (B in Table 10) consisted in replacing the forcing data of interest, i.e., either boundary or initial conditions, by forcing data from an equivalent season that is selected based on the criteria that it would force the output variable in the opposite tercile than the observed one.

Experiment 2B: In the case of the deterministic experiment to investigate the impact of boundary conditions on discharge, for example Autumn 2000, the SEAS5 forcing data covering the target season, i.e., September to November 2000, is replaced by ERA5 forcing data from an Autumn season (September to November) where tp falls into the “below” tercile, e.g., Autumn 1997. ERA5 forcing data for the boundary conditions are specifically selected in the opposite tercile as the “above” observed tercile for discharge in Autumn 2000 to challenge the forecasting model and observe whether the skill remains.

Experiments 3-4B: Similar experiment replacing the forcing data over the warm-up period (3B) and over the warm-up period and transition month (4B) were performed to investigate the impact of initial conditions on a specific forecasted variable.

Experiment 1A: The application of deterministic experiments requires to identify to which forcing variables each output variable is the most sensitive to. This was assessed through an additional preliminary experiment where the ERA5 forcing data for a specific input variable, e.g., tas, covering the target season, i.e., September to November 2000, is replaced by tas-ERA5 forcing data from a randomly-selected Autumn (September to November, e.g., Autumn 2005). Experiment 1A has been performed 25 times for each of the input variables of the hydrologic and lake models in order to draw a complete sensitivity analysis.

Table 10: List of experiments performed and case study sites which have completed each experiment to-date.

Data to be replaced	A - Probabilistic	B - Deterministic
1 – One specific variable in boundary conditions – SEAS5 over target season	Norway, Spain	Not applicable
2 – All variables in boundary conditions – SEAS5 over target season	Norway, Spain	Norway (average)
3 - Initial conditions – ERA5 over warm-up period	Norway, Spain	Norway (average), Germany
4 - Initial conditions – ERA5 over warm-up and SEAS5 over transition month	Norway, Spain	Germany

The outputs of the probabilistic experiments were used to produce tercile plots, calculate ROCSS and FairRpss scores, while the outputs of the deterministic experiments, devoid of probability, were used to calculate goodness of fit metrics, i.e., correlation coefficient, NSE, Bias and RMSE.

3 Results & Discussion

3.1 Skill of the seasonal climate and impact model forecasts

SEAS5 lake: The comparison of Table 2 and Table 9 is already useful to identify possible transfer of forecasting skills from the SEAS5 seasonal climate forecasts to the impact models. Table 2 shows that only 0 to 10% of the SEAS5 climate forecasts are skilful. However, for some case-studies, e.g., Norway and Australia, there is a higher number of skilful forecasts for spring than for the other seasons. Interestingly, spring also stands out in the skilful forecasts for the impact models in Norway (Table 9). For Australia, there is only one combination of season/variable/tercile which is skilful and happens to be in spring, but it could well be a spurious result (Table 9). For Spain and Germany, such a clear connection between SEAS5 climate forecasts and impact model outputs is not as straightforward. We can thus hypothesize that the skills of impact model forecasts in Norway is more inherited from the SEAS5 data than at other case studies. In contrast, skills of the impact model forecasts at the other case-studies is hypothesized to originate from the legacy of the warm-up period or from the parametrization of the inflow-outflow water balance.

SEAS5 fish: Regarding the fish forecasting model, given the complete absence of skill in the SEAS5 climate forecasts (Table 2), the only skilful forecast of the impact model is thought to originate from the warm-up period (i.e. explanatory variables in the model which describe the fish preparedness for migration, and how the discharge forecast is influenced by conditions prior to the target season).

SiNaï: Climate forecasts were not included as predictors in the models used to forecast lake water chemistry/ecology in Norway, as these variables proved to be relatively insensitive to seasonal weather (see Section 1.5). The basis for the predictions of the next season was therefore observed water quality the previous season, and interactions between chemical and algal variables. The skill therefore relies heavily on temporal autocorrelation.

3.2 Datasets and error profiling

Most of the studies involving seasonal climate predictions have embraced the perfect-model assumption (Wood et al., 2016). When assessing model predictability, forecasts are compared to pseudo-observations, i.e., model simulations driven with weather observations, and not to observations. When using model simulations, the model error related to model structural error and parameter uncertainty is not accounted for. Hence, the predictability is likely to be overestimated.

On the other hand, observed time series are rarely perfectly continuous and often include data gaps, shifts in sampling methods over time or other limitations. The absence of a coherent time series covering sufficiently long periods often prevents their use for the evaluation of forecasts in modelling studies. In fact, if the observed time series are not long enough, the probabilistic skills of the forecasts can't be fully assessed. There is thus a clear challenge related to the incorporation of observations in the analysis of model-based predictability. While the ideal practice is to compare forecasts to observations, quality observations are often lacking, leaving only pseudo-observations as validation datasets.

Below, we first review the observations that meet the coherence and length criteria before comparing pseudo-observations to observations as well as forecasts to pseudo-observations and observations, when possible. This assessment provides some insights on SEAS5_lake forecast skill and how model error affects them.

Table 3 provides an overview of the data available at each case-study as well as their sampling frequency. For each SEAS5_lake, the corresponding observations have been plotted in supplementary figures in Appendix 2 (Fig. S1, S3, S5, S7). Fig. S2, S4, S6 and S8 also show, for each season (winter – DJF; spring – MAM; summer – JJA; autumn – SON), the number of days for which an observation is available for each of the climate and impact variables.

For climate observations, most of the variables used for modelling in Vansjø have complete timeseries with daily observations, except for wind (uas and vas) and long-wave radiation. For Sau, complete observations are available for all variables but only for a period of nine years. Similarly, for Wupper, complete observations for all variables are available from 2003. These constraints in time and for certain variables have prevented the use of meteorological observations to force process-based models and have motivated our choice to force our models with the ERA5 data to produce the pseudo-observations. Regarding impact variables, discharge is the only variable for which daily observations are available at all case-studies throughout the whole or significant parts of the modelled time period. For Wupper, coherent water temperature datasets are available for both surface and bottom temperature throughout the whole period. For Sau, a long record of monthly surface temperature is also available from 1997, while bottom temperature data are more sporadic and more frequent after 2010. For Vansjø, observations for surface and bottom water temperature are only available from 2005 to 2015 over May to October. Hence, due to data scarcity, the observations for bottom temperature in Sau, the observations for surface and bottom temperature in Mt Bold, as well as surface and bottom temperature for Vansjø in all seasons except summer were not considered.

3.2.1 Pseudo-observations vs observations

Table 11 shows some statistics describing how well pseudo-observations (ERA5_lake) represent real observations for several impact variables and for each season. While Tables 4 and 5 displayed quite good statistics for year-round time series, Table 11 shows that there are significant discrepancies between seasons.

For Vansjø, we clearly see that discharge is usually well-simulated by ERA5_lake, except in Summer where the NSE is negative and the other statistics mostly show a decrease in modelling skill. Regarding water temperature, given that observations mostly cover the summer season (May to October), the statistics reported for summer are similar to those reported for the all-year in Tables 4 and 5. Note however that despite the absence of observations over November to April, ERA5_lake captures quite well the timing of ice-on and to a lesser degree that of ice-off. Upon integration of water temperature observations in winter and spring, the skill of the model could be even better regarding ice-off.

For Sau, discharge is consistently well simulated, although a little decrease in modelling skill can be seen in Summer. The modelling skills related to surface temperature, on the other hand, show a strong seasonality as winter and summer are much better simulated than spring and autumn. For Wupper, discharge is best simulated in autumn and winter while summer shows less skilful simulations. Regarding water temperature, the model shows strong skills for both surface and bottom in winter and significantly lower skills for the other seasons, especially summer for bottom temperature.

3.2.2 Forecasts vs pseudo-observations or real observations

Some of the forecasting skills which are shown to be significant when making the "perfect-model" assumption (SEAS5_lake vs ERA5_lake) might not be significant when they are assessed against observations. On the other hand, some of the forecasting skills might not be significant when assessed against ERA5_lake pseudo-observations, although they would have been if assessed against observations. Below, we describe how well forecast skill, as assessed against pseudo-observations or observations, coincide, and discuss the likely causes for discrepancies. Table 11 includes ROCSS values calculated for SEAS5_lake against ERA5_lake (ROCSS_ERA5; same values as ROCSS in Tables 13-16) but also against observations (ROCSS_obs) when possible.

For Vansjø, while significant forecast skills were reported for discharge in spring when assessed against ERA5_lake, the assessment against observations of discharge showed that none of the terciles or seasons are associated with a skilful forecast. Nevertheless, we can note that the ROCSS_obs for the lower and upper terciles in Spring for discharge, stand out as being more positive compared to the rest of the ROCSS_obs values. Regarding water temperature, winter and spring showed the highest number of skilful forecasts (Table 9). Unfortunately, water temperature observations are not covering enough time of the winter and spring seasons to be included in a forecast skill assessment. However, ice observations are available providing a completely independent validation dataset. In fact, ROCSS_ERA5 and ROCSS_obs values agree quite well regarding ice-off which provide further confidence that water temperature (at least for surface temperature) forecasts are indeed skilful in winter and spring. Lastly, ROCSS_ERA5 and ROCSS_obs reported for water temperature in summer both illustrate the absence of forecasting skill for these variables, although the ROCSS_obs value for the upper tercile for summer bottom temperature is high. The latter result should be taken with caution since ROCSS_obs values have been derived from a 10-year observation record only, in contrast to 25 years for ROCSS_ERA5.

Preliminary results from the Wupper reservoir in Germany are also shown in Table 11. Here, the observed data for discharge covers the entire time span at daily resolution for the same period as the hindcast, whereas the water temperature is at biweekly resolution for the same period. Results show that:

- *Discharge*: The forecast appeared to perform better when compared to observations rather than pseudo-observations. In autumn, for example, the lower tercile has a significant ROCSS, and the same for the summer average tercile. The number of data

points is the same for the observations used when calculating ROCSS_obs and ROCSS_ERA5, so it does seem that there are more windows of opportunity when we use real observations for discharge.

- *Water temperature*: for spring and autumn ROCSS_obs are worse than ROCSS_ERA5. For winter and summer, it is the opposite and ROCSS_obs are higher. For the above average tercile in summer there is even significant skill, which is promising as it is associated with summer heat waves which are of great interest. For bottom temperature, only in autumn is the skill worse for ROCSS_obs compared to ROCSS_ERA5, otherwise it is better for all other seasons and terciles, with some significant ROCSS_obs and substantially more so than ROCSS_ERA5. However, due to the big difference in the number of data points between real observations and Pseudo- observations, it is not possible to claim that using real observations will produce a better and more skilful forecast

In theory ROCSS_ERA5 should always be higher than ROCSS_obs, as ROCSS_ERA5 makes a perfect model assumption (i.e. does not include uncertainties introduced by the catchment and lake models, which certainly exist). It is therefore very curious that ROCSS_obs was often improved at the German site. As yet, only Vansjø and Wupper have completed this assessment, but this work is planned at the other sites for the forthcoming paper planned based on this work.

Table 11: Goodness of fit statistics for pseudo-observations SEASONAL MEANS – ERA5_lake, comparing ERA5_lake to observations, as well as comparison of the ROCSS for SEAS5_lake forecasts calculated against ERA5_lake (pseudo-observations) and observations.

			NS	R^2	RMSE	RMSE/sd	bias	ROCSS_ERA5			ROCSS_obs		
								lower	middle	upper	lower	middle	upper
Norway	Discharge	WI	0.90	0.94	2.0	0.31	0.4	0.08	0.23	-0.22	0.11	-0.2	-0.07
		SP	0.72	0.80	2.0	0.52	-1.0	**0.58***	0.13	**0.54***	0.36	0.13	0.34
		SU	-0.82	0.61	3.5	1.32	2.7	-0.28	-0.04	-0.55	-0.04	-0.19	-0.5
		AU	0.60	0.91	4.4	0.62	-3.2	0.08	0.23	-0.22	0.07	0.12	0.08
	Surface Temp.	SU	0.55	0.70	0.4	0.64	0.2	0.33	0.39	-0.13	-0.43	-0.29	-0.54
	Bottom Temp.	SU	0.79	0.88	0.8	0.43	0.4	-0.12	0.14	-0.35	0.1	0.25	0.76*[a]
	Ice-on		0.97	0.99	2.2	0.16	1.8	[b]					
	Ice-off		0.36	0.76	19.3	1.09	-14.7	**0.69***	0.29	**0.75***	**0.55***	0.25	**0.68***
Spain	Discharge	WI	0.88	0.89	3.9	0.34	-0.6	0.11	0.18	0.08	0.40	**0.52***	0.08
		SP	0.58	0.74	5.5	0.63	-3.2	0.01	-0.22	0.33	0.37	-0.11	0.33
		SU	0.51	0.62	3.5	0.69	-1.6	0.40	-0.01	0.27	**0.73***	-0.02	0.23
		AU	0.73	0.74	4.0	0.51	-0.8	0.01	0.39	0.40	0.17	0.46	**0.47***
	Surface Temperature	WI	0.54	0.62	0.7	0.64	0.2	0.32	-0.66	-0.15	0.45	-0.27	**0.62***
		SP	-0.74	0.17	1.3	1.27	-0.8	0.20	0.18	0.19	0.12	-0.10	0.42
		SU	0.12	0.40	1.1	0.87	-0.6	0.42	-0.13	**0.57***	-0.25	0.03	-0.08
		AU	-1.28	0.18	2.0	1.46	-1.3	0.22	0.11	0.47	**0.79***	0.08	0.39
	Bottom Temperature	WI						-0.02	0.54	0.27	0.13	-0.25	0.38
		SP						0.44	0.12	**0.86***	-0.10	0.10	0.33
		SU						**0.53***	0.36	**0.72***	0.42	0.31	**0.63***
		AU						**0.5***	0.55	**0.64***	0.25	-0.60	-0.33
Germany	Discharge	WI	0.85	0.87	1.1	0.38	0.0	-0.28	0.17	-0.34	-0.17	0.21	0.18
		SP	0.21	0.69	1.4	0.87	-1.1	-0.47	-0.03	-0.34	-0.22	-0.04	-0.33
		SU	0.08	0.40	1.0	0.94	-0.5	-0.59	-0.13	-0.34	-0.07	**0.58***	0.04
		AU	0.52	0.69	1.7	0.68	1.0	-0.29	-0.33	0.03	**0.53***	0.3	-0.09
	Surface Temperature	WI	0.60	0.92	0.5	0.59	-0.4	-0.22	0.04	-0.09	0.23	0.08	0.18
		SP	-0.06	0.40	1.0	1.01	-0.2	0.2	-0.14	0.38	0.09	-0.12	-0.04
		SU	-1.89	0.56	1.5	1.66	1.3	-0.25	-0.18	-0.22	0.2	0.06	**0.52***
		AU	-2.97	0.62	1.6	1.95	-1.5	0.23	0.42	0.3	0.18	-0.2	-0.1
	Bottom Temperature	WI	0.87	0.87	0.3	0.33	0.0	-0.16	-0.3	0.27	-0.06	0.03	0.23
		SP	-5.01	0.49	1.2	2.40	1.0	0.38	-0.43	0.32	0.41	0.04	**0.46***
		SU	-8.63	0.26	3.8	3.04	3.6	0.27	-0.03	**0.52***	**0.51***	0.21	**0.49***
		AU	-0.16	0.33	0.7	1.05	-0.1	0.27	0.26	0.15	0.11	0.4	-0.07
Australia	Discharge	WI	-0.75	0.16	1.23	1.26	-0.88						
		SP	0.40	0.51	0.58	0.74	-0.23						
		SU	0.26	0.64	0.95	0.82	0.34						
		AU	-0.67	0.41	1.61	1.23	-1.27						

[a]The upper tercile for summer bottom temperature at Vansjø is only assessed against a 10-year record.
[b]Ice-on typically occurs between November and December which is the boundary that we chose to delimit autumn and winter. This prevented obtaining ROCSS values for ice-on.

3.2.3 Other statistics as a complementary assessment of forecasting skill

The fair Ranked Probability Skill Score (FRPSS) is also a common measure to evaluate probabilistic forecasts. The Ranked probability Score (RPS) measures the reliability of the probabilities given by the forecast compared to the distribution of observations among categories (Above, Normal, Below), and it is calculated as a squared-error between the probability distribution of forecasts and observations across categories. The FRPSS compares RPS for the forecast with RPS using average climate conditions as a predictor, i.e., it is a measure of the relative improvement of the probability forecast over climatology in predicting the category that the observations fell into. An FRPSS of 1 indicates perfect multicategory probabilistic forecasts, while a zero value or less indicates the forecast is not superior to the reference climatology. In contrast to the ROCSS, the FRPSS takes into account the bias between the forecast and the observations/pseudo-observations.

At our sites, we see that FRPSS was sensitive to the amplitude of the variation in forecast value *within* a season. E.g. water temperature forecasts in Vansjø for spring and autumn encompass a much larger temperature range than forecasts for winter and summer, and have correspondingly large FRPSS (Table 12). When we remove the within-season trend from the forecasted water temperature, then we see that ROCSS tends to increase slightly, whilst FRPSS tends to decrease in spring and autumn (Table 12).

Table 12: Impact of the seasonal amplitude (or intra-seasonal trend) on ROCSS and FRPSS in Vansjø. ROCSS and FRPSS are calculated from normal daily timeseries (Normal) and from homogenized timeseries (detrend). The homogenized time series have been corrected for the difference in monthly means.

Season	Variable	ROCSS (* = significant at 95% level)						FRPSS	
		Below		Normal		Above			
		Normal	*Detrend*	Normal	*Detrend*	Normal	*Detrend*	Normal	*Detrend*
	Surface T	**0.48***	**0.51***	-0.12	0.13	-0.17	**0.63***	0.08	0.17
	Bottom T	**0.48***	**0.63***	0.01	0.21	**0.53***	0.32	0.10	0.19
Spring	Q	**0.58***	**0.58***	0.13	0.13	**0.54***	**0.54***	0.09	0.09
	Surface T	**0.75***	**0.84***	0.14	*0.41*	**0.53***	**0.77***	**0.66***	0.47
	Bottom T	**0.56***	**0.81***	0.05	**0.61***	**0.68***	**0.77***	**0.74***	0.44
Autumn	Surface T	-0.52	0.29	-0.24	0.30	-0.12	0.06	**0.74***	0.44
	Bottom T	-0.53	0.27	0.03	0.22	-0.26	-0.08	**0.74***	0.42

As well as ROCSS, we calculated goodness-of-fit of simulated and observed in terms of the R^2 and RMSE, and averaged these over ensemble members. Results for climate variables are shown in Table 13. We can see that the distribution of mean R^2 agrees broadly with the ROCSS-based results discussed earlier in Section 3.2.2, in terms of the seasons which show best model performance. For example, seasonal climate forecasts appear to be most reliable in spring in Norway compared to other seasons and case studies.

Table 13: Standard statistics for climate variables among the various seasons and case studies. All SEAS5 climate variables associated with a mean R^2 (n = 25) higher than 0.05 with respect to ERA5 data are highlighted with a cross. The average R^2 for all variables, as well as the standardized RMSE (RMSE*) are also shown. RMSE is standardized by variable and season meaning that RMSE* below 1 display a relatively smaller error for a given season compare to the other seasons.

		Variables										
		psl	rlds	rsds	tas	tcc	tdps	tp	uas	vas	R^2	RMSE*
Norway	WI							X	X		0.05	1.25
	SP	X	X	X	X	X	X	X	X	X	0.07	0.95
	SU			X	X			X	X	X	0.05	0.81
	AU	X	X							X	0.05	0.99
		hurs	tp	rsds	tas	tas_min	tas_max	wss			R^2	RMSE*
Spain	WI	X	X		X	X	X				0.06	0.96
	SP	X			X		X				0.04	1.16
	SU				X						0.04	0.98
	AU	X	X					X			0.05	0.89
		hurs	tp	rsds	tas	cc	wind				R^2	RMSE*
Germany	WI		X				X				0.05	0.95
	SP	X	X		X						0.05	1.16
	SU										0.04	0.98
	AU					X					0.05	0.92
		psl	rlds	rsds	tas	tcc	tdps	tp	uas	vas	wss	R^2	RMSE*
Australia	WI	X	X	X	X	X	X				X	0.08	0.89
	SP		X				X		X	X		0.05	1.10
	SU	X									X	0.05	0.99
	AU	X	X	X					X	X	X	0.06	1.02

Results for impact model forecasts are shown in Table 14, and again we see that the distribution of mean R2 agrees with results based on ROCSS in Section 3.2.2: SEAS5_lake correlates better with ERA5_lake for the variables and seasons with significant ROCSS values in each case study. This shows that even these “simple” deterministic summary statistics can yield information on forecasting skill, that agrees with and complements results obtained using probabilistic skill scores.

Table 14: Standard statistics for impact variables among the various seasons and case studies. All SEAS5_lake variables associated with a mean R^2 (n = 25) higher than 0.05 with respect to ERA5_lake data are highlighted with a cross. The average R^2 for all variables, as well as the standardized RMSE (RMSE*) are also shown. RMSE is standardized by variable and season meaning that RMSE* below 1 display a relatively smaller error for a given season compare to the other seasons.

Site	Season	Variables				
		Surface Temp.	Bottom Temp.	Discharge	R^2	RMSE*
Norway	WI	X			0.05	0.71
	SP	X	X	X	0.20	0.95
	SU			X	0.05	1.29
	AU			X	0.05	1.04
Spain	WI				0.03	1.05
	SP	X	X		0.21	0.83
	SU	X	X		0.18	0.87
	AU		X	X	0.09	1.24
Germany	WI		X	X	0.05	1.15
	SP	X	X		0.11	0.82
	SU		X		0.14	0.98
	AU				0.03	1.05
Australia	WI	X	X		0.12	0.62
	SP	X	X	X	0.06	0.82
	SU	X	X		0.08	1.64
	AU		X		0.06	0.92

3.3 Synergies between seasonal climate forecast skill and impact model sensitivity

Given the relatively higher number of windows of opportunity reported for the SEAS5_lake workflow in Spain and Norway compared to the other sites, only these two case-studies are considered in the assessment of the sources of predictability below. The source of predictability for Si/Naï is discussed after, while no assessment was performed for SEAS5_lake in Germany and SEAS5_fish in Ireland, since very few windows of opportunity were found in these cases. Indeed, at Ireland in particular, where only one window of opportunity was found, it could be a false positive (given the 5% significance level used to search for windows of opportunity), so investigations into the origin of the skill in here could be misleading.

Experiment 1A aimed to determine which climate variable each modelled impact variable is sensitive to. This experiment consisted in replacing the data for a specific input climate variable by random data and quantify the deterioration of the modelled outputs compared to default outputs (ERA5_lake). Table 11 summarizes the influence of input variables on impact variables. The R^2 and mean raw bias were calculated by comparing experiment 1A outputs to ERA5_lake. Lower R^2 values represent more influence of the input variables on the model outputs. Results for Norway and Spain are that:

- *Norway:* Several of the most influential input variables are also associated with some forecasting skill, as shown in Table 2. Indeed, in spring, tas, uas and vas are among the most influential input variables for discharge, surface and bottom temperature and are reported to have some forecasting skill (Table 2). In winter, only tcc falls into these categories, however, it is only the fourth most influential input variable and for bottom temperature only. We thus hypothesize there will be more forecasting skills transferred

from SEAS5 to the impact model forecasts in spring than in winter, and that most of the forecasting skills in winter are a legacy of the warm-up period.

- *Spain*: Impact variables here were relatively less sensitive to climate forcing compared to Norway, although similar variables were key for water temperature (e.g., tas, rsds, uas/vas), as in Norway. Lower sensitivity to weather variables could be due to lower amplitude changes both within and between seasons or differences in site characteristics (e.g. Sau is a very actively managed reservoir, which experiences large water level fluctuations that uncouple water level and from incoming discharge).

Table 15: Influence of input variables on impact variables. The mean R^2 and mean raw bias were calculated by comparing experiment 1A outputs to ERA5_lake for 25 independent runs. Only impact variables associated with some forecast skills are shown here.

Site	Seasons (months)	Impact Variable	Input variable replaced by random data	Mean R^2	Mean raw bias
Norway	Winter (12,1,2)	Surface Temperature	tas	0.33	-0.18°C
			uas/vas	0.84	-0.14°C
			tdps	0.75	-0.14°C
		Bottom Temperature	tas	0.21	-0.12°C
			uas/vas	0.56	-0.14°C
			tdps	0.62	-0.11°C
			tcc	0.88	+0.02°C
	Spring (3,4,5)	Discharge	tp	0.16	+1.3%
			tas	0.37	+2.0%
		Surface Temperature	tas	0.92	-0.20°C
		Bottom Temperature	tas	0.84	-0.16°C
			uas/vas	0.86	-0.09°C
			rsds	0.94	+0.09°C
Spain	Spring (3,4,5)	Bottom Temperature	tas	0.86	+0.02°C
	Summer (6,7,8)	Surface Temperature	tas	0.85	-0.06°C
			rsds	0.90	-0.04°C
			uas/vas	0.94	-0.07°C
		Bottom Temperature	tas	0.96	+0.02°C
			uas/vas	0.96	+0.02°C
			rsds	0.96	+0.02°C
	Autumn (9, 10, 11)	Bottom Temperature	-		

3.4 Inheritance of forecasting skills

The results of experiment series A (probabilistic) and B (deterministic) were used to investigate the influence of the boundary conditions, the initial conditions as well as the transition month on the forecasting skills. When many season/variable/tercile combination showed some forecasting skill (Table 2), we used the probabilistic approach (e.g., exp. 2A), i.e., for Norway and Spain. In the other cases, we used a deterministic approach (e.g., exp. 2B). Note that these results are preliminary and the probabilistic experiments should be run more than five times to provide a full probabilistic overview. Also, note that experiments 2A and 2B excluded the SEAS5 transition month which is included in the experiments 4A and 4B.

3.4.1 Skill due to SEAS5 (boundary conditions; experiment 2A)

Norway: ROCSS values reported for experiment 2A show some variability compared to the default SEAS5_lake outputs (Table 16). Only two of the skilful season/variable/tercile combinations, both in spring, show consistently similar or larger ROCSS values for the experiment 2A than for SEAS5_lake, i.e., surface temperature in the lower tercile, and discharge in the upper tercile (Table 16). The forecasting skill of these two season/variable/tercile combinations is thus expected to come from another source than the SEAS5 boundary conditions over the target season. In the other cases, there is no clear signal that SEAS5 boundary conditions influence the forecasting skill except for bottom temperature in the lower tercile in spring and for bottom temperature in the upper tercile in winter. Indeed, in these two cases, the average ROCSS values are decreased by more than 0.05 compared to SEAS5_lake. Note that for three skilful season/variable/tercile combinations, the introduction of random boundary conditions yielded an increase in the average ROCSS value. This suggests that in these cases, the use of SEAS5 climate forecasts deteriorates the skill inherited from the initial conditions.

Spain: Experiment 2A has been conducted in Spain, but results have not been aggregated yet for incorporation into Table 16.

Table 16: Comparison of ROCSS from SEAS5_lake (N) and experiment 2A outputs (n=5; R1–5) for upper and lower terciles. The ROCSS values for experiment 2A falling within ± 0.05 of that for SEAS5_lake are shown in green; those above or below these thresholds are reported in blue and orange, respectively. * are significant positive ROCSS.

Site	Season		ROCSS (* = significant at 95% level)													
		Tercile	Lower							Upper						
		Variable	N	*R1*	*R2*	*R3*	*R4*	*R5*	*Avg.*	N	*R1*	*R2*	*R3*	*R4*	*R5*	*Avg.*
Norway	Win	Surface Temp	**0.48***	*0.39*	*0.43*	***0.63****	***0.77****	***0.47****	0.54	-0.17						
		Bottom Temp	**0.48***	***0.60****	*0.43*	*0.28*	***0.50****	*0.43*	0.45	**0.53***	*0.35*	*0.38*	*0.32*	***0.64****	***0.51****	*0.44*
	Spr	Discharge	**0.58***	***0.86****	***0.70****	***0.89****	*0.47*	*0.43*	0.67	**0.54***	***0.77****	***0.62****	***0.70****	***0.57****	***0.56****	*0.64*
		Surface Temp	**0.75***	***0.70****	***0.73****	***0.70****	***0.73****	***0.77****	0.73	**0.53***	***0.56****	***0.58****	***0.68****	*0.48*	*0.36*	*0.53*
		Bottom Temp	**0.56***	***0.62****	***0.52****	*0.42*	*0.34*	*0.41*	0.46	**0.68***	***0.62****	***0.69****	***0.63****	***0.60****	***0.66****	*0.64*
Spain	Spr	Bottom Temp.	0.44							**0.86***						
	Sum	Surface Temp.	0.42							**0.57***						
		Bottom Temp.	**0.53***							**0.72***						
	Aut	Bottom Temp.	**0.50***							**0.64***						

3.4.2 Skill due to initial conditions (experiment 3A)

Norway: The ROCSS values reported for experiment 3A show less variability compared to the experiment 2A outputs above, and more consistently show a decrease in forecasting skill, emphasizing the importance of the initial conditions (Table 17). Only two of the skilful season/variable/tercile combinations, both in spring, show similar or better forecasting skills with random initial conditions, i.e., bottom temperature in the lower tercile, and surface temperature in the upper tercile (Table 17). Interestingly, the average ROCSS value for bottom temperature in the lower tercile in spring is even larger than that of the default SEAS5_lake. This finding is consistent with the loss of skill in experiment 2A related to the boundary conditions as described above. Hence, apart from bottom temperature in the lower tercile in spring, most of the forecasting skills seem to originate from the initial conditions.

Spain: Experiment 3A has been completed for Spain, but have not yet been aggregated for inclusion in Table 17.

Table 17: Comparison of ROCSS from experiment 3A outputs (n =5; W1–5) and SEAS5_lake (N). The ROCSS values for experiment 3A falling within ± 0.05 of that forSEAS5_lake are shown in green, those above or below these thresholds are reported in blue and orange, respectively.

Site	Season		ROCSS (* = significant at 95% level)													
		Terciles	Lower							Upper						
		Variables	N	*R1*	*R2*	*R3*	*R4*	*R5*	*Avg.*	N	*R1*	*R2*	*R3*	*R4*	*R5*	*Avg.*
Norway	Win.	Surface Temp	**0.48***	*0.31*	*0.42*	*0.32*	*0.22*	*0.34*	0.32	-0.17						
		Bottom Temp	**0.48***	*0.18*	*0.25*	*0.36*	*0.04*	*0.22*	0.21	**0.53***	*0.38*	*0.36*	*0.36*	*0.38*	*0.34*	*0.36*
	Spr.	Discharge	**0.58***	***0.49****	***0.60****	*0.12*	*0.41*	*0.23*	0.23	**0.54***	***0.49****	*0.33*	*0.05*	*0.21*	*0.28*	*0.27*
		Surface Temp	**0.75***	***0.71****	*0.40*	***0.57****	***0.71****	***0.69****	0.62	**0.53***	*0.38*	***0.59****	***0.55****	*0.45*	***0.44****	*0.48*
		Bottom Temp	**0.56***	***0.55****	***0.54****	***0.62****	***0.60****	***0.80****	0.62	**0.68***	***0.66****	***0.54****	***0.59****	***0.60****	***0.65****	*0.61*
Spain	Spr.	Bottom Temp.	0.44							**0.86***						
	Sum.	Surface Temp.	0.42							**0.57***						
		Bottom Temp.	**0.53***							**0.72***						
	Aut.	Bottom Temp.	**0.50***							**0.64***						

3.4.3 Skill due to initial conditions and transition month (experiment 4A)

Norway: The ROCSS values reported for experiment 4A show even less variability than the experiment 3A above, and an even more consistent decrease in forecasting skill, emphasizing the combined importance of the initial conditions and the transition month (Table 14). All of the skilful season/variable/tercile combinations have lost more than 0.05 units in their ROCSS values. The fact that all average ROCSS values for experiment 4A are smaller than those for experiment 3A shows that significant forecasting skill is also inherited from the SEAS5 transition month. Note again that the average ROCSS value for bottom temperature in the lower tercile in spring is the one that has decreased the least, only by 0.06 units, emphasizing the importance of boundary conditions in this case.

Spain: Experiment 3A has been completed for Spain, but have not yet been aggregated for inclusion in Table 17.

Table 18: Comparison of ROCSS from experiment 4A outputs (n =5; W1–5) and SEAS5_lake (N). The ROCSS values for experiment 4A falling within ± 0.05 of that for SEAS5_lake are shown in green, those above or below these thresholds are reported in blue and orange, respectively.

Site	Season		ROCSS (* = significant at 95% level)													
		Terciles	Lower							Upper						
		Variables	N	*R1*	*R2*	*R3*	*R4*	*R5*	*Avg.*	N	*R1*	*R2*	*R3*	*R4*	*R5*	*Avg.*
Norway	Win.	Surface Temp	**0.48***	*-0.03*	*0.23*	*0.19*	*0.27*	*0.07*	0.15	-0.17						
		Bottom Temp	**0.48***	*-0.05*	***0.53****	*-0.06*	***0.48****	*-0.03*	0.17	**0.53***	***0.45****	*0.28*	*0.39*	***0.62****	*-0.03*	*0.34*
	Spr.	Discharge	**0.58***	***0.62****	*0.05*	*0.38*	*0.38*	*0.11*	0.31	**0.54***	*0.20*	*0.11*	*0.06*	*0.32*	*0.18*	*0.17*
		Surface Temp	**0.75***	*0.31*	*0.47*	*0.37*	***0.58****	***0.69****	0.48	**0.53***	*0.26*	***0.58****	*0.24*	*0.4*	*0.03*	*0.30*
		Bottom Temp	**0.56***	***0.52****	*0.44*	***0.46****	*0.42*	***0.61****	0.49	**0.68***	***0.62****	***0.54****	***0.57****	***0.44****	*0.37*	*0.51*
Spain	Spr.	Bottom Temp.	0.44							**0.86***	*0.29*	*0.29*	*0.23*	*0.70*	*0.29*	*0.23*
	Sum.	Surface Temp.	0.42							**0.57***	*0.04*	*0.04*	*0.23*	*-0.03*	*0.18*	*0.11*
		Bottom Temp.	**0.53***	*0.02*	*-0.08*	*0.25*	*0.04*	*0.04*	0.08	**0.72***	*0.12*	*0.12*	*0.12*	*0.68*	*0.34*	*0.25*
	Aut.	Bottom Temp.	**0.50***	*-0.05*	*-0.30*	*-0.05*	*-0.30*	*-0.30*	-0.06	**0.64***	*0.30*	*0.30*	*0.30*	*0.30*	*0.03*	*0.17*

3.5 Repartition of forecasting skills among initial, transition and boundary conditions

By comparing the ROCSS values obtained through the various experiments, we can map where forecasting skill is coming from (Table 19). For Norway, we can see that the majority (on average 30% of the skill, but ranging up to 60%) is coming from the warm-up period, i.e. much of the skill is due to knowledge and representation of initial conditions in the catchment

and lake, and inertia which means this influence carries through to the target season. However, substantial skill is also coming from the SEAS5 transition month (mean of 19% across windows in Norway). SEAS5 for the target season adds little value apart from perhaps to bottom temperature in winter and spring.

Results from Spain will be summarized in future work.

Table 19: Contribution of the initial conditions, transition and warm-up conditions to final forecasting skills. The contributions are calculated by subtracting the various ROCSS values from experiments 2A, 3A and 4A to the original ROCSS from SEAS5_lake. Negative contributions indicate that it worsens the skill.

Site	Season	Variable	Tercile	Source of the forecasting skill (contribution to ROCSS value)					
				SEAS5 (season)	SEAS5 (transition)	SEAS5 (all)	ERA5 (warm-up)	? interactions	ROCSS original
Norway	Winter	Surface Temperature	lower	-0.06	0.18	0.12	0.16	0.20	**0.48***
		Bottom Temperature	lower	0.03	0.04	0.07	0.27	0.14	**0.48***
			upper	0.09	0.02	0.11	0.17	0.25	**0.53***
	Spring	Discharge	lower	-0.09	0.06	-0.03	0.35	0.26	**0.58***
			upper	-0.10	0.10	0.00	0.27	0.27	**0.54***
		Surface Temperature	lower	0.02	0.13	0.15	0.13	0.47	**0.75***
			upper	0.00	0.18	0.18	0.05	0.30	**0.53***
		Bottom Temperature	lower	0.10	0.13	0.23	-0.06	0.39	**0.56***
			upper	0.04	0.10	0.14	0.07	0.47	**0.68***

Table 20: Contribution in percentage of the initial conditions, transition and warm-up conditions to final forecasting skills. The contributions are calculated by subtracting the various ROCSS values from experiments 2A, 3A and 4A to the original ROCSS from SEAS5_lake. Negative contributions indicate that it worsens the skill.

Site	Season	Variable	Tercile	Source of the forecasting skill (%; contribution to ROCSS value)					
				SEAS5 (season)	SEAS5 (transition)	SEAS5 (all)	ERA5 (warm-up)	? interactions	ROCSS original
Norway	Winter	Surface Temperature	lower	-13	37	24	33	43	100
		Bottom Temperature	lower	6	8	14	56	30	100
			upper	17	4	21	32	47	100
	Spring	Discharge	lower	-16	11	-5	60	44	100
			upper	-19	18	0	50	50	100
		Surface Temperature	lower	3	18	20	17	62	100
			upper	0	34	34	9	57	100
		Bottom Temperature	lower	18	24	41	-11	69	100
			upper	6	15	21	10	69	100

3.6 Results of deterministic experiments

These results are preliminary, and rely on a previous set-up of the workflow in the German site (which has since been improved, such that there are now more windows of opportunity for bottom water temperature).

In the German case study, we had skilful predictions only during summer season for bottom temperature for the above-average tercile. The below-average tercile showed positive skill, but it was below the significance threshold. The Wupper reservoir is dimictic and typically thermally stratified during the summer season. The bottom temperature is mainly related to the thermal conditions in the lake in the previous mixing season, as well as being affected by

the water withdrawal via the bottom outlet. The original ROCSS value for the upper tercile of SEAS5_lake outputs was 0.52* and for the lower tercile 0.27, with an FRPSS of 0.51.

The workflow is designed to start from January first, two years before the target season. For example, to forecast the summer of 2003, we start from 2001-01-01. Such a period was required as a spin-up time of the catchment model and the lake model.

The main idea of these experiments is to trace the source of predictability and inspect the effects of initial conditions and boundary conditions on the target season's predictability, based on our knowledge of the anomaly that we get from the seasonal forecast. We achieve this by challenging the target season with opposite conditions in the lake in the previous season (spin-up time) and examine whether we still can have the same skills entirely or partially, as well as challenging the target season with opposite boundary conditions during the target season. The principle is to evaluate the forecasting system by forcing it with a warm-up period where the relevant meteorological variable of the previous season of the boundary conditions has an opposite anomaly (see Fig 7). In this case, the most relevant meteorological variable is the air temperature. In other words, where we see skill in the warmer than average bottom temperature tercile, we use a year that has a colder than average air temperature during warm up. We created the synthetic warmup data from ERA5 time series. The criteria were, choosing years whose antecedent season to the target season has an opposite probability tercile compared to the target season in terms of air temperature. In this example, we chose years where air temperature was below-average (dots in the plots Fig. 8). We chose the years of 1995, 1996, 2004, 2006, 2010, 2013, 2015 and 2016. And to test boundary conditions we chose the years 1993, 1996, 1998, 2000, 2004, 2011, 2012 and 2014 which had pseudo-observations from ERA5 where air temperature was below-average during the target season.

Experiment 2B was designed as follows to change the boundary conditions of the target seasons:

- create an ensemble of 8 members that represent the colder years in summer from ERA5 data. For the months (May, June, July, August). The rest of the meteorological variables remained untouched.
- This synthetic forecast ensemble was appended to the original warm-up data from ERA5 to keep the initial conditions of the forecast similar to the original simulations
- we replicated the synthetic data ensemble over every single year from 1993-2016

The setup of experiment 3B was:

- Find years whose time-series ended with a colder spring.
- Copying all the meteorological variables from January two years before.
- Replicating this warm-up data before each summer season from every year.
- Repeating the same steps for all other chosen years for the synthetic warm-up.

For example, to use the conditions of spring 2004 as initial conditions, we copied the time series from January 2002 until the end of April 2004, then replicate this time-series prior to each summer from 1993-2016 while accounting for leap years. This method will ensure that the skills of predicting higher than average anomalies are challenged from a warm-up perspective.

Experiment 4B is based on the same motivation and method of 3B. However, in this experiment, we changed the lead month (transitional period) that precedes the target season. Using the same example, we copied the time series from January 2002 until the end of May

2004. This experiment aims to recognize the effect of the transitional month on the skills of prediction during the target season.

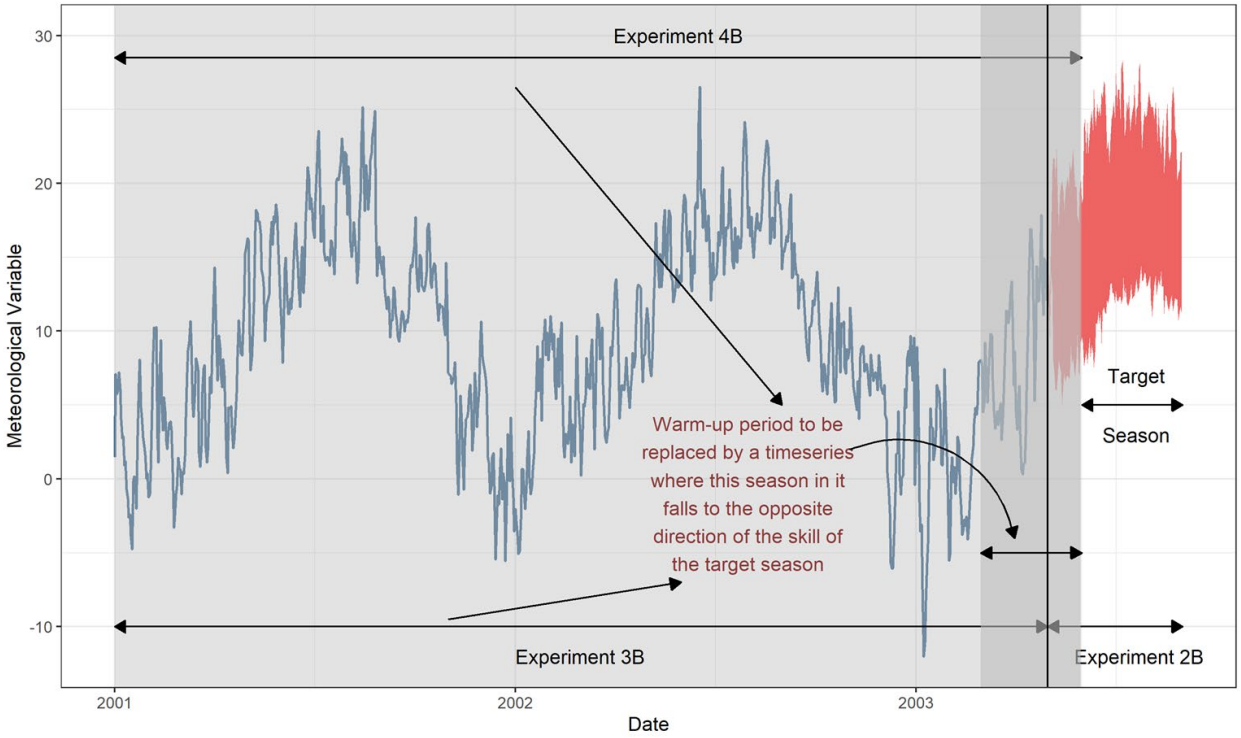

Figure 7: Schematic of the setup of experiments 3B and 4B. Showing the periods that are to be replaced in respect to the target season.

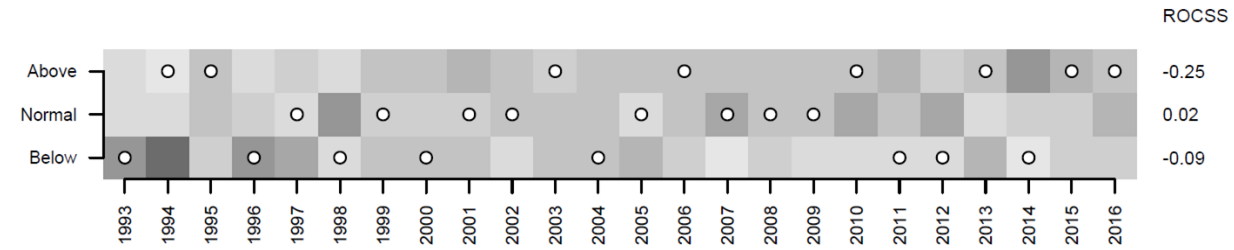

Figure 8: Summer air temperature (June, July, August).

The results of experiment 2B showed ROCSS value for the upper tercile of SEAS5_lake outputs was **-0.15** and for the lower tercile **0.23**. The effect of forcing colder conditions on the target season had a major effect by removing all the skill from our predictions (Fig. 9). For example, the year 2003 was a major heatwave that shows high probability for above average temperature in the original simulations, and when forced with colder conditions, the anomaly for this year has totally disappeared.

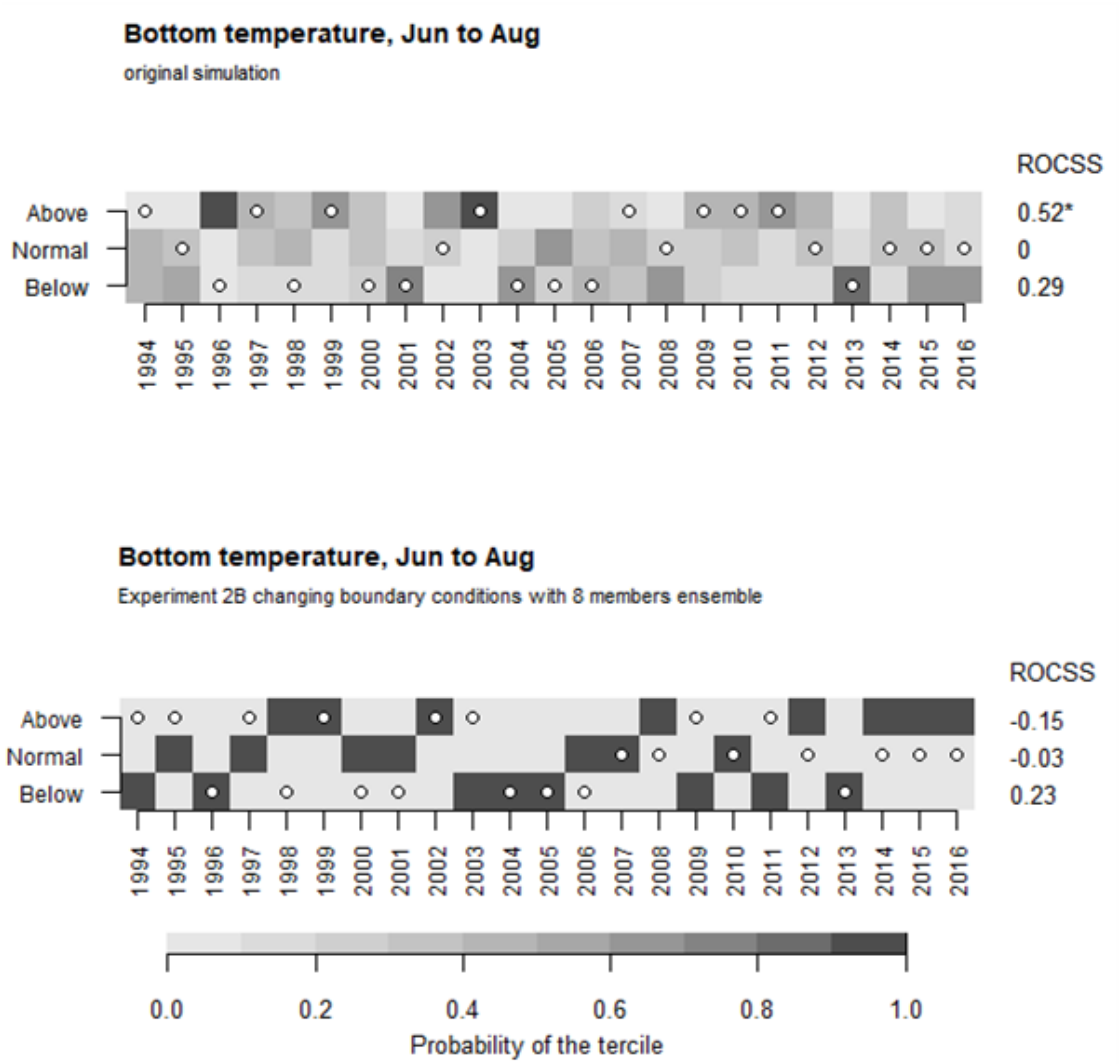

Figure 9: Standard tercile plots of the bottom temperature for summer (June, July, August) for the original simulation without any changes and for experiment 2B forced by a synthetic ensemble of 8 members chosen from cold ERA5 summers and repeated every year.

The ROCSS values reported for the experiment 3B (Table 21) show some variability compared to the default SEAS5_lake outputs (Table 12). While in some cases the skill of the below average tercile has increased due to the colder warm-up, only one of the skilful variable/tercile combinations in one year, 2013, showed significant skill similar to the original ROCSS. The skills showed consistent positive values for ROCSS ranging from 0.37 to 0.53, with the exception of the year 1995 which had remarkably lower ROCSS value (Table 21). This suggests that in these cases, the warm-up period of ERA5 has an important impact on the seasonal forecast in the target season. FRPSS were slightly lower compared to the real conditions. All this indicates that the forecast skill of the lake model contains a systematic overestimation arising from the warm-up periods as the water body's temperature has a specific memory effect. This is particularly valid for the bottom temperatures.

Table 21: Experiment 3B: Impact of the warm-up period on forecasted bottom temperature ROCSS in the Wupper reservoir. The original ROCSS value for the above average tercile of SEAS5_lake outputs was 0.52* and for below than average was 0.27.

Warm up year	Tercile	ROCSS (* = significant at 95% level)	FRPSS
1995	above	0.14	0.47
	below	0.23	
1996	above	0.43	0.5
	below	0.38	
2004	above	0.3	0.5
	below	0.39	
2006	above	0.49	0.48
	below	0.24	
2010	above	0.46	0.5
	below	0.33	
2013	above	**0.53***	0.48
	below	0.23	
2015	above	0.44	0.49
	below	0.32	
2016	above	0.37	0.49
	below	0.31	

The ROCSS values reported for the experiment 4B (Table 22) show similar variability and trends compared to the experiment 3B outputs above, and more consistently show an increase in forecasting skill for the below average tercile, emphasizing the importance of the initial conditions and transitional period. No skilful variable/tercile combinations were recorded for both above or below average terciles. In experiment 3B, the simulations using warm-up data from the year 2013 showed a significant skill for the upper tercile, however this skill disappeared in experiment 4B. Except for the year 2016, generally, the skill in terms of ROCSS decreased. The loss of skill in experiment 4B compared to 3B elucidate the essential impact of the transition month on the forecast. In a similar trend, the FRPSS is lower in 4B compared to 3B.

Table 22: Experiment 4B: Impact of the warm-up period and the transitional (lead) month on bottom temperature forecasts for the Wupper reservoir. The original ROCSS value for the above average tercile of SEAS5_lake outputs was 0.52* and for below average was 0.27.

Warm-up year	Variable	Tercile	ROCSS (* is significant at 95% level)	FRPSS
1995	bottom temperature	above	0.11	0.4
		below	0.08	
1996	bottom temperature	above	0.34	0.43
		below	0.18	
2004	bottom temperature	above	0.37	0.45
		below	0.27	
2006	bottom temperature	above	0.42	0.35
		below	0.12	
2010	bottom temperature	above	0.44	0.48
		below	0.25	
2013	bottom temperature	above	0.23	0.45
		below	0.24	
2015	bottom temperature	above	0.29	0.45
		below	0.32	
2016	bottom temperature	above	0.41	0.45
		below	0.31	

We can conclude that the skill seems to originate from a combination of 1) the initial conditions for summer predictions, especially during the transitional period; and 2) mainly the boundary conditions during the target season. Moreover, since the forcing data during the target seasons substantially influenced the seasonal prediction in the lake temperature, we can see a connection between the climate forcing and the impact variable. Thus, the skilful upper tercile for bottom temperature in summer appears to represent a genuine window of opportunity for predictions.

4 Concluding remarks

We have carried out a number of experiments to help explore the sources of skill in seasonal forecasting at the case study sites. Results show that much of the skill is due to the initial conditions and system inertia. However, SEAS5 does impart skill during the transition month in particular, and in Norway in particular adds some value to predictions particularly during spring.

Part II: Visualisation and clear communication of uncertainty

Seasonal forecasts are probabilistic and uncertain, and so for them to be useful to decision making the quality of the forecasts must be clearly communicated with an informed end user base. To this end, we held a workshop on visualising and communicating uncertainty at the 3rd WATExR project meeting in Ireland in 2019. This workshop was used to set the scene for the kinds of quality information that should be communicated in forecasting tools and ensure that all case study sites included the necessary information, as well as to discuss with end users their preferences for how the information should be presented.

In summary, two main sources of quality information must be communicated with forecasts:

1. 'Predictability' of the future state of the environment: this is how well the different ensemble members e.g. of SEAS5 agree with one another
2. Historic skill information: how well the forecast performed during the past, when compared to observations.

In advance of the workshop, various methods for communicating these two kinds of quality information were gathered into a handout, together with the information included in the draft forecasting tools at each site. We then talked through general recommendations from the EUPORIAS project, an FP7 project which had a particular focus on communicating forecast quality. General recommendations which were passed on to WATExR developers included:

- There is no "one visualisation fits all" solution, so work together with end users
- Don't provide forecasts when the forecasting system has no skill, as research has shown that end users tend to be influenced by the forecast regardless of the lack of skill (and they shouldn't be, as it has no value).
- Provide qualitative skill and uncertainty categories (e.g. None, Low, Medium, High) and visual cues (e.g. colour, opacity) to help users "make sense" of skill information. Take care with colour (e.g. is red for dry, or for skill? And take colour blindness into account).
- Any attempt to classify skill as 'good' or 'low' is subjective, and will vary by sector, so decide on thresholds together with users.
- If users simply require quantile likelihoods with a measure of skill, then consider using a tabular format.
- Consider a single measure of skill which combines both the probability of the class AND the historic skill of the system.
- A tiered/layered approach may be useful to avoid clutter/confusion, where different levels of information may be selected by different user groups

The preliminary forecasting layouts produced at each of the case study sites were examined in relation to these recommendations, and we highlighted where improvements could be made:

Case study	Skill information present		No forecast when skill too low?	Qualitative skill categories?	Tiered approach relevant?
	Reliability	Historic skill			
Australia	✔	✔	✔	✔?	✔
Denmark	✔	✗			
Germany	✔	✗			
Ireland	✔	✔	✔	✔?	✔
Norway	✔	✔	✗	✔	
Spain	?	✔	✗		
Sweden					

We then had a feedback task, where end users were provided with a forecast for a select season for their study site, together with forecast quality information. They were then asked a number of questions relating to their interpretation of the forecast and its reliability/confidence, and ideas/preferences for improvements that could be made:

1. What would be your understanding of the take home message for this event?
    - What will the weather be like? What might happen in the lake?
    - How trustworthy are these projections?
2. Is enough information provided for you to assess forecast uncertainty & trustworthiness?
3. Do you like how uncertainty is presented, or would you prefer something different? Simpler/more info?
4. Would you find qualitative categories useful for summarising forecast uncertainty (if not already used)?
5. What level of historic skill would you want before taking action based on a forecast? E.g. what level of skill should be considered ‘Poor’, ‘Moderate’ or ‘Good’?
6. What level of forecast reliability (agreement between model forecasts, e.g. probability of a given tercile) would you want before acting based on predictions? How should we define thresholds for e.g. ‘Poor’, ‘Moderate’, ‘Good’ reliability?

Concrete answers were not captured from each site for each question, but rather the questions were used to help frame more general discussions on quantifying and visualising management-relevant uncertainty. In general, there was a range of preferences in how forecast quality should be presented, from a simple text summary to interest in full details of how skill was derived. Results of these discussions for each site will be used to guide further development and refinement of the different forecasting tools.

References

Coron, L., Thirel, G., Delaigue, O., Perrin, C., & Andréassian, V. (2017). The suite of lumped GR hydrological models in an R package. *Environmental Modelling & Software*, *94*, 166–171. https://doi.org/10.1016/j.envsoft.2017.05.002

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., … Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hipsey, M. R., Bruce, L. C., Boon, C., Busch, B., Carey, C. C., Hamilton, D. P., Hanson, P. C., Read, J. S., de Sousa, E., Weber, M., & Winslow, L. A. (2019). A General Lake Model (GLM 3.0) for linking with high-frequency sensor data from the Global Lake Ecological Observatory Network (GLEON). *Geoscientific Model Development*, *12*(1), 473–523. https://doi.org/10.5194/gmd-12-473-2019

Jackson-Blake, L. A., Sample, J. E., Wade, A. J., Helliwell, R. C., & Skeffington, R. A. (2017). Are our dynamic water quality models too complex? A comparison of a new parsimonious phosphorus model, SimplyP, and INCA-P. *Water Resources Research*, *53*(7), 5382–5399. https://doi.org/10.1002/2016WR020132

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., Keeley, S. P. E., Mogensen, K., Zuo, H., & Monge-Sanz, B. M. (2019). SEAS5: The new ECMWF seasonal forecast system. *Geoscientific Model Development*, *12*(3), 1087–1117. https://doi.org/10.5194/gmd-12-1087-2019

Mercado-Bettín, D., Clayer, F., Shikhani, M., Moore, T.N., Frías, M.D., Jackson-Blake, L.A., Sample, J., Iturbide, M., Herrera, S., French, A.S., Norling, M.D., Rinke, K., Marcé, R. (2021). Forecasting water temperature in lakes and reservoirs using seasonal climate prediction. *Water Research*, *201*.

Wood, A. W., Hopson, T., Newman, A., Brekke, L., Arnold, J., & Clark, M. (2016). Quantifying streamflow forecast skill elasticity to initial condition and climate prediction skill. *Journal of Hydrometeorology*, *17*(2), 651-668.

Appendix 1: Empirical fish model workflow

1. Hydrologic modelling

Fish migration is highly dependent on river flows, and so hydrologic modelling was carried out to provide input to a statistical fish migration model.

Water levels for the largest freshwater lake in the Burrishoole catchment, Lough Feeagh, have been recorded by the Irish Environmental Protection Agency (EPA) since 1976 (Irish EPA, 2020). Daily discharge was estimated from water levels using a rating curve (Fig 1). This curve was established using 20 paired discharge – water level measurements from 2016 - 2018 for the combined outflows of the two short channels that link Lough Feeagh to the tidal lagoon, Lough Furance (Kelly *et al.*, 2020). The R package *bbmle::mle2* (Bolker & R Core Team, 2017) was used to fit the rating curve.

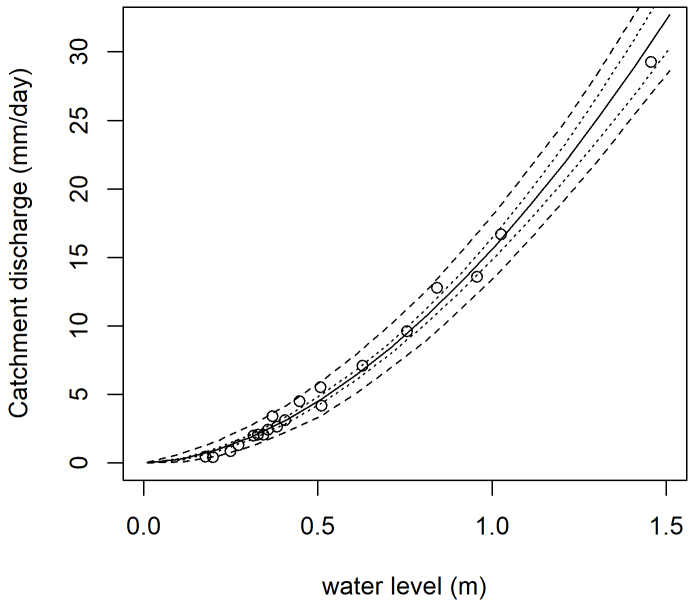

Figure 1: A rating curve illustrating modelled relationship between Lough Feeagh lake level and the combined outflows from the Burrishoole catchment.

Catchment outflow discharge was modelled using the GR4J lumped rainfall-runoff model, using the *airGR* R package (Coron et al., 2020). To calibrate GR4J, discharge data for the combined outflows of the Mill Race and Salmon Leap channels were used. The model was driven by bias-corrected ERA5 precipitation and estimated daily potential evapotranspiration. GR4J was calibrated using Michel's procedure provided in *airGR*, whereby the modified Kling-Gupta Efficiency (KGE) was applied to Box-Cox transformed flows. The Box-Cox transformation adds weight to low flow conditions (Santos, Thirel & Perrin, 2018).

2. Water temperature modelling

Lough Feeagh water temperature profiles (2 minute resolution) have been recorded by an Automatic Water Quality Monitoring Station (AWQMS) since 2004 (de Eyto *et al.*, 2019), and river temperature of the Mill Race (which is effectively Lough Feeagh surface temperature) have been recorded since 1960 (Dillane *et al.*, 2018).

We used 2m depth Lough Feeagh data from 2004 to 2014 and *bbmle::mle2* (Bolker & R Core Team, 2017) to estimate plausible parameter values of a four-parameter air temperature to water temperature statistical model, whereby we assumed daily water temperature was linearly correlated with lagged mean air temperature. Our model extended the three-parameter lagged air to river temperature model proposed by Ducharne (2008), by allowing lag period to shorten during the stratification period of Lough Feeagh (reflecting multi-year average Schmidt stability dynamics for May to September; Calderó-Pascual *et al.* (2020)), thus adjusting for the shallower mixed layer depth during stratification and reduced volume responsive to meteorological forcing (Piccolroaz, Toffolon & Majone, 2013).

3. Fish count modelling

The aim was to demonstrate a transferable seasonal forecasting workflow. To this end, the goal was to identify one out of the many potential correlative models for each species, for which simulated counts were sufficiently similar to observations to be useful for further evaluation of seasonal forecasting. To this end, we used glmmTMB owing to its flexibility for modelling dispersion and structural and sampling zeros (Brooks et al., 2017), and we used DHARMa (Hartig, 2019) for statistical validation.

For each species, daily counts were our response variable, and we derived predictor variables from publicly available non-climate data (i.e., lunar and solar data from suncalc; Thieurmel & Elmarhraoui (2019)) in addition to air temperature and precipitation from ERA5 reanalysis. We derived daily values for moonlight exposure, photoperiod, potential evapotranspiration (using the Hargreaves-Samani equation implemented in drought4R; Bedia & Iturbide (2019)), catchment discharge (Section 1), photoperiod-weighted degree days and linear rates of change in air temperature and photoperiod.

Our modelling was necessarily exploratory; however, we followed a short series of steps to specify a model for each species and we limited complexity of each model to aid interpretability in the context of alternative applications of seasonal forecasting to diadromous fish migrations. Firstly, for all species, we specified a “full model” set of four environmental covariates (discharge; water temperature; moonlight exposure; and a proxy for physiological preparedness for migration (and change in water temperature for eels)) and all possible three-way interactions for salmon and trout and two-way interactions for eel. These covariates provided the starting point for a (conditional effects formula only) model structure with Poisson error distribution and random intercept year. We subsequently adjusted the model structure (in accordance with DHARMa’s statistical validation plots) with the addition of quadratic terms in the first instance, followed by substitutions of error distributions, specifications of dispersion formulas, and inclusion of zero-inflation formulas if necessary. This model adjustment procedure does not constitute a formal statistical modelling workflow; it simply represents one of multiple possible paths to obtaining a statistically valid predictive model using only our selected covariates.

To maintain an out-of-sample validation for our seasonal forecasting protocol, we used a leave-one-year-out strategy to fit a model for each species and each year of the dataset. Prior to fitting (training), each covariate was scaled to standard deviation units. For subsequent tests of predictive performance on held-out data, we used scaling parameters derived from the training set using *caret::preProcess* (Jed Wing et al., 2019). We checked for consistency between statistical model predictions and observations by inspecting simulated (quantile) residuals using DHARMa (Hartig, 2019). We carried out two sets of checks for each species based on residuals calculated for models fitted using data from 1980 to 1992 (for initial model specification) and 1993 to 2019 (for a test set held out from the original fitting procedure).

Appendix 2: Visualisation of the datasets

The following figures show the complete datasets for ERA5_Lake, SEAS5_Lake and the forcing timeseries for the various climate variables. The titles of the figures or the legend are mainly self-explanatory. Here are a few explanations in case some clarifications are needed:

Climate variables:

tas – Daily average 2-m air temperature (°C)
tas_min – Daily minimum 2-m air temperature (°C)
tas_max – Daily minimum 2-m air temperature (°C)
tp – Daily total precipitation (mm) (°C)
tdps – 2 metre dewpoint temperature
tcc (or cc) – Cloud cover (unitless from 0 to 1)
psl – Surface pressure (Pa)
hurs – Near surface humidity (%)
rsds – Surface Downwelling Shortwave Radiation (W/m^2)
rlds – Surface Downwelling Longwave Radiation (W/m^2)
uas – West-east wind (m/s)
vas – South-north wind (m/s)
wss (or wind) – wind speed (m/s)
petH – potential evaporation (mm)

Impact variables:

Q: Daily mean discharge (m^3/s)
temp_0m/temp_surf/surftemp: Daily mean water temperature at 0 m depth (°C)
temp_15m/temp_bottom/bottemp: Daily mean water temperature at 15 m depth or the lake bottom (°C)

Fig S1: Vansjø – hydrological and lake modelling dataset

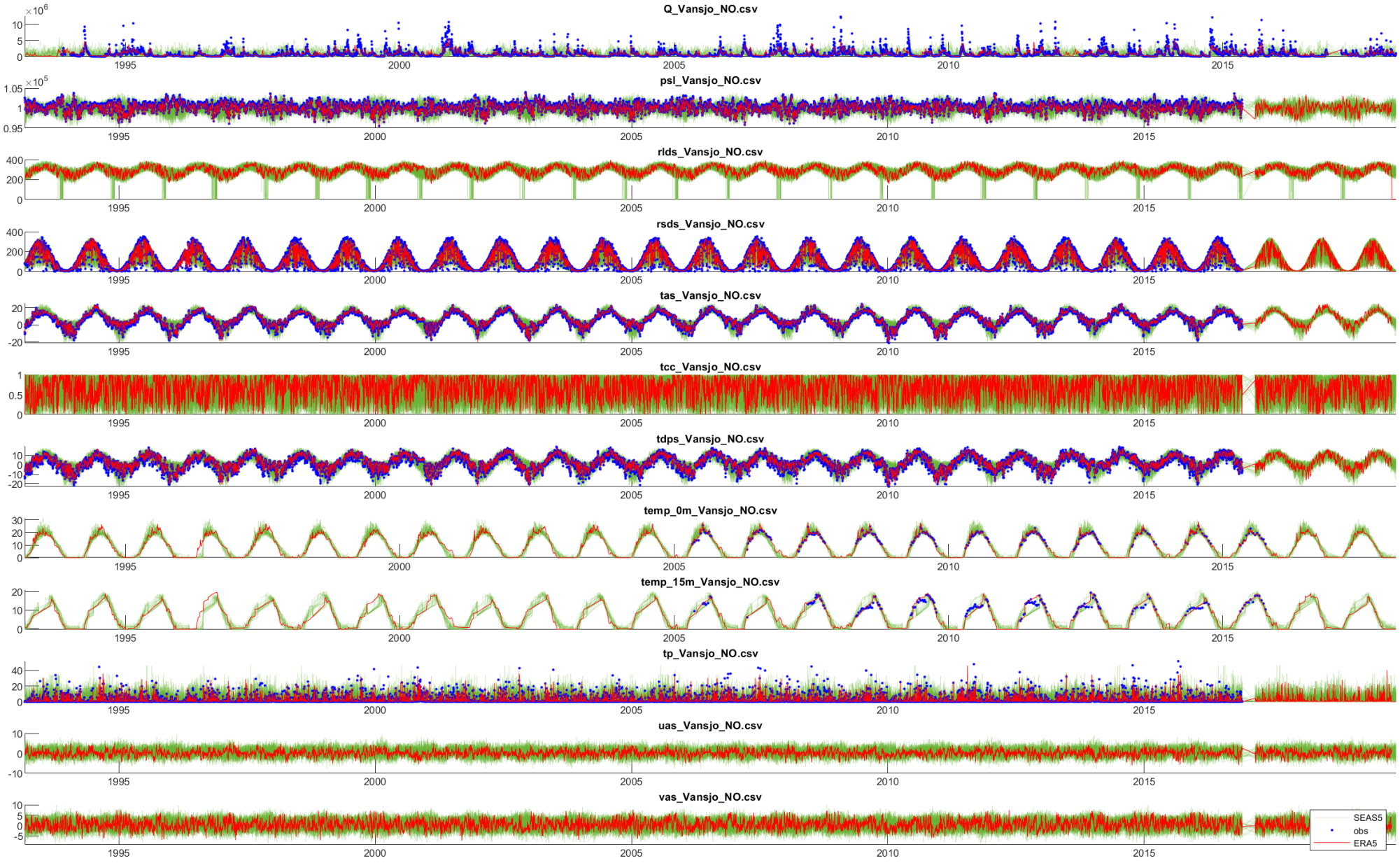

Fig. S2: Vansjø – Data coverage for hydrological and lake modelling, observations versus ERA5

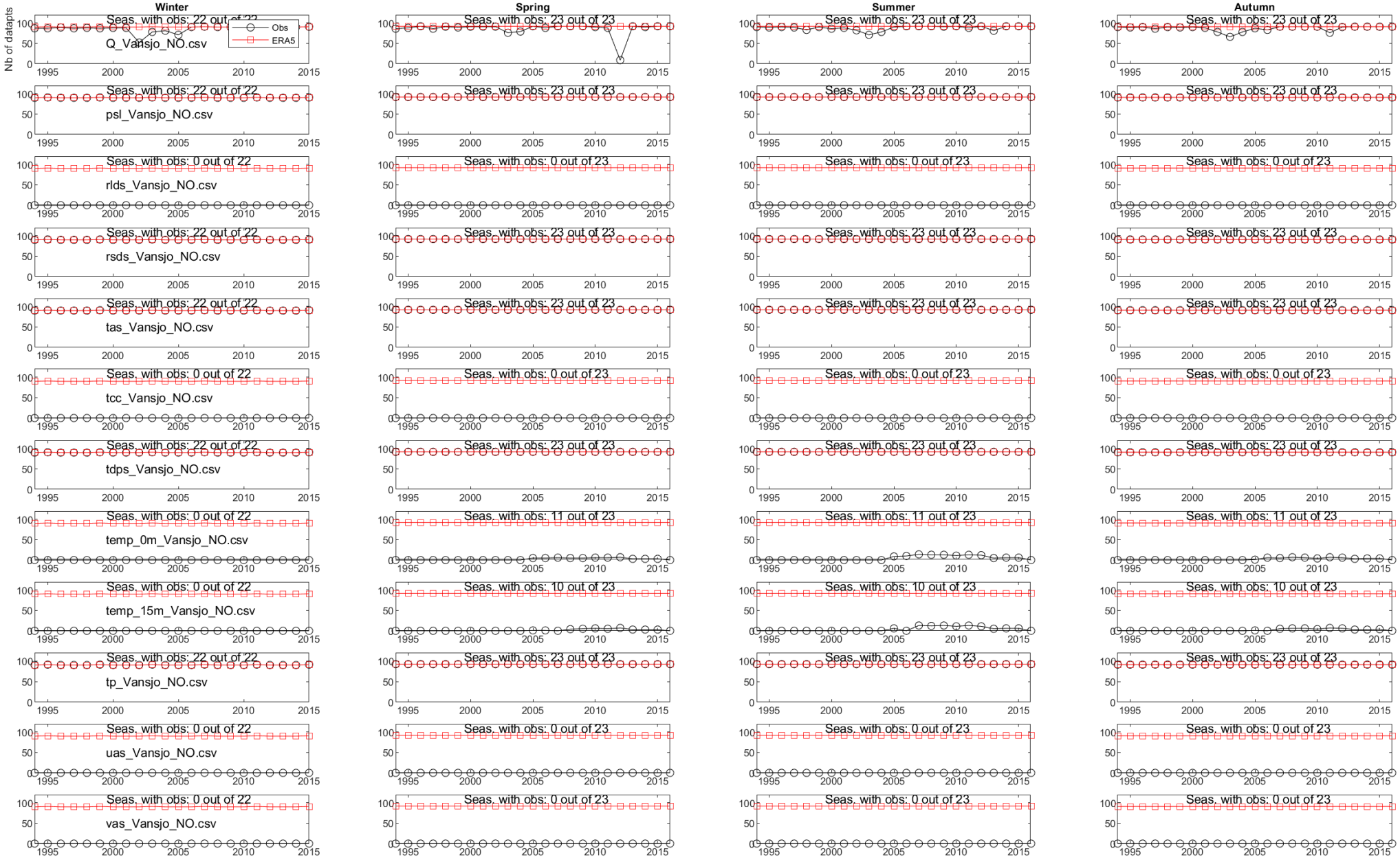

Fig S3: Sau – hydrological and lake modelling dataset

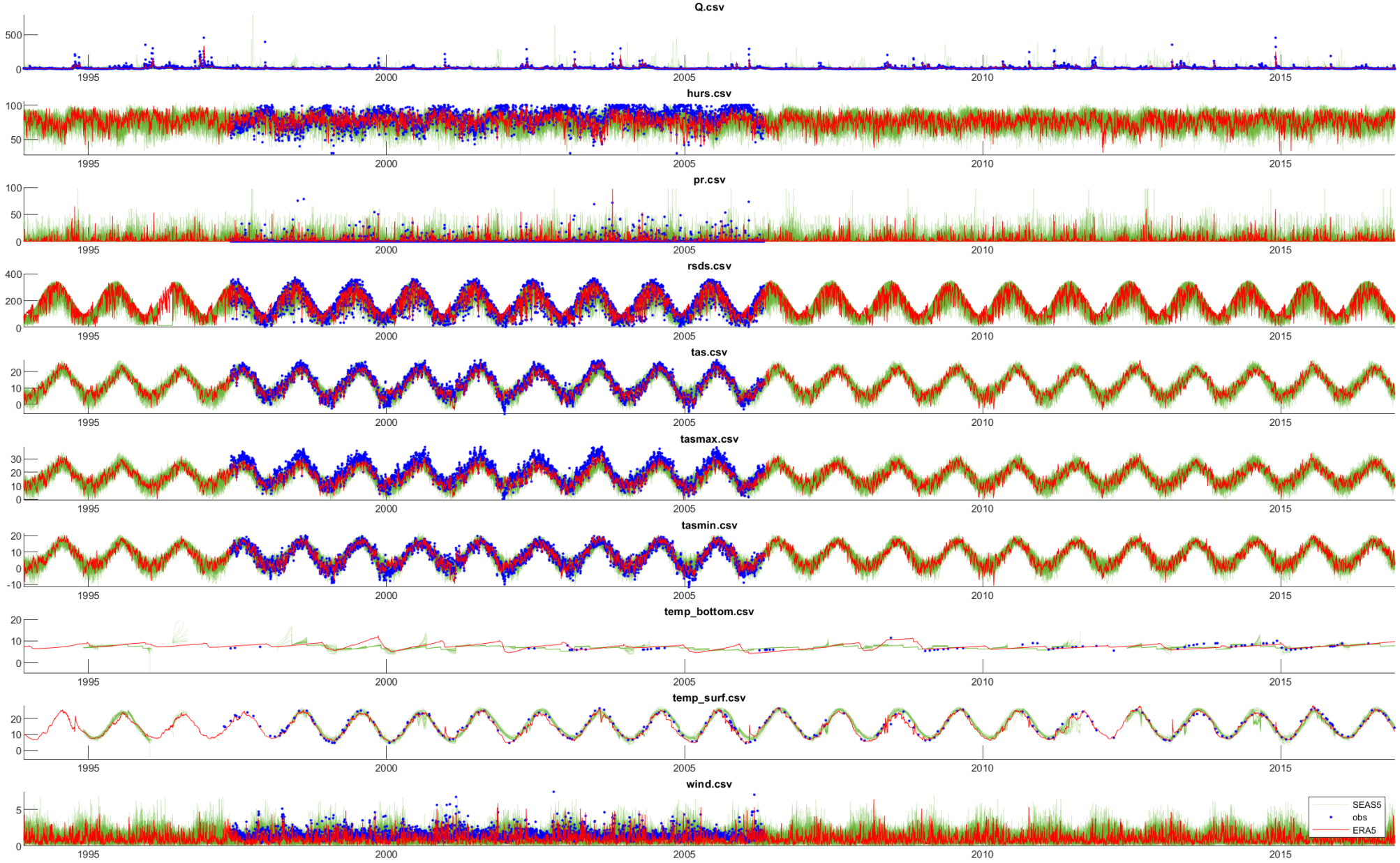

Fig S4: Sau – Data coverage for hydrological and lake modelling, observations versus ERA5

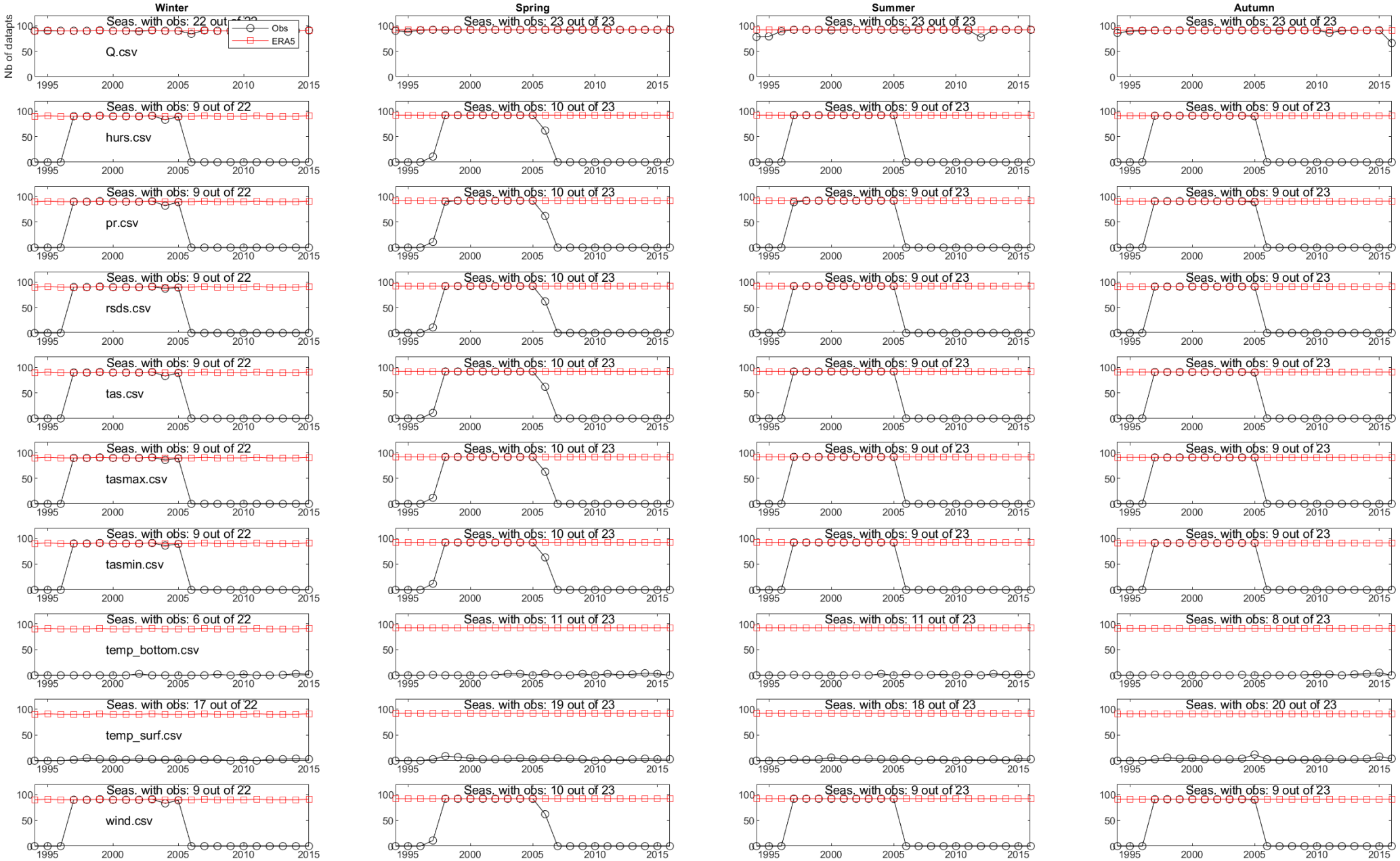

Fig S5: Wupper – hydrological and lake modelling dataset

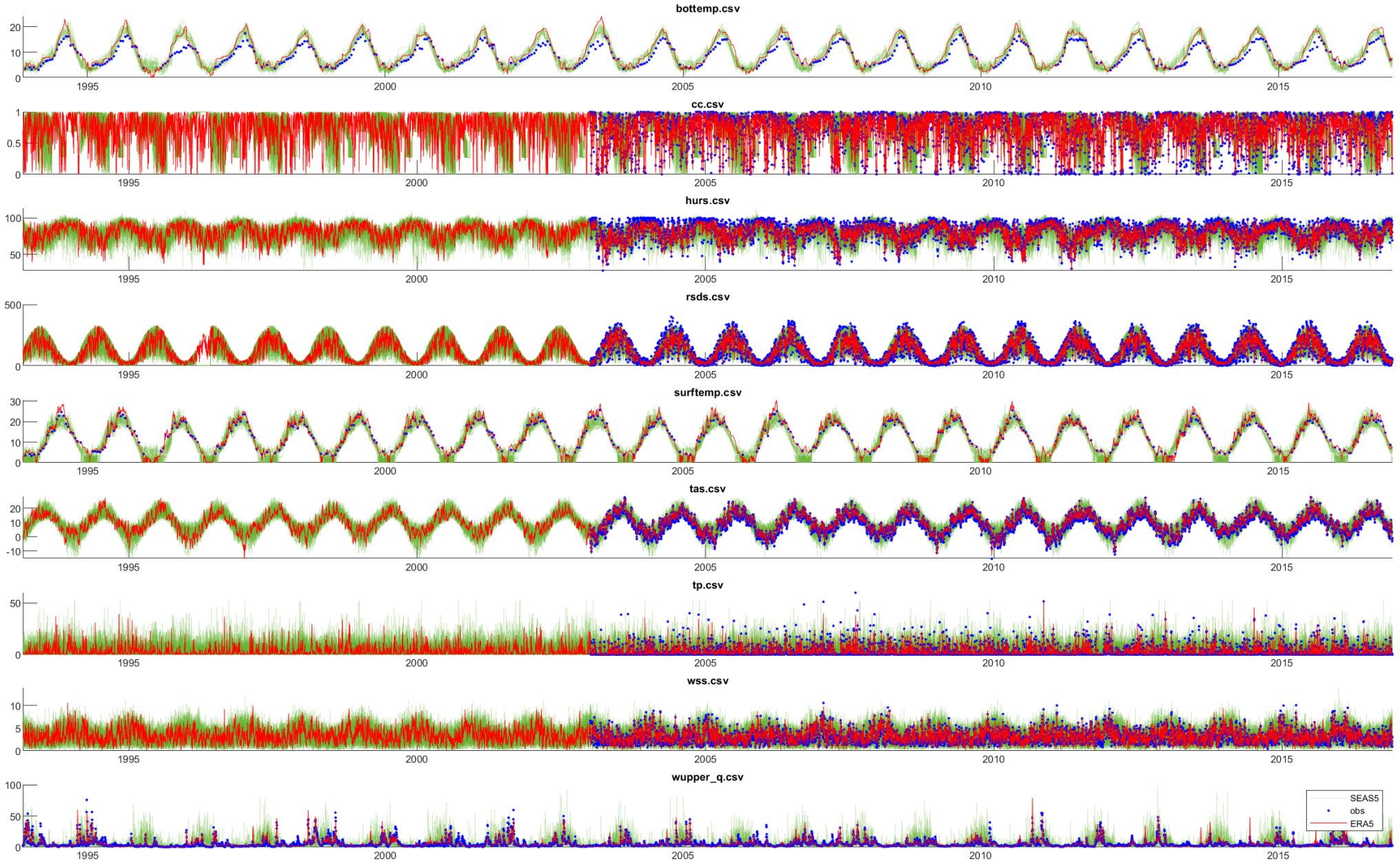

Fig S6: Wupper – Data coverage for hydrological and lake modelling, observations versus ERA5

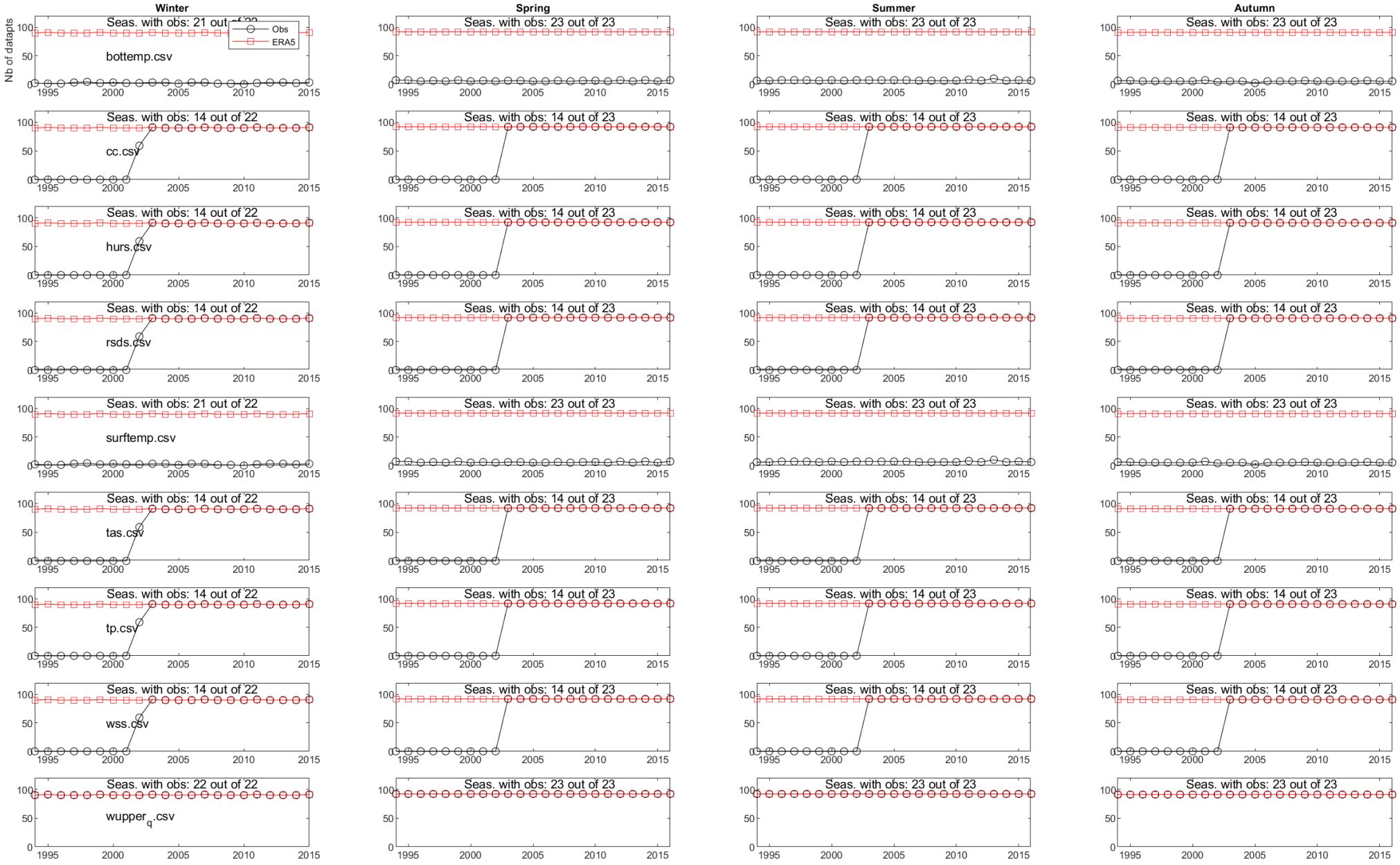

Fig S7: Mt Bold – hydrological and lake modelling dataset

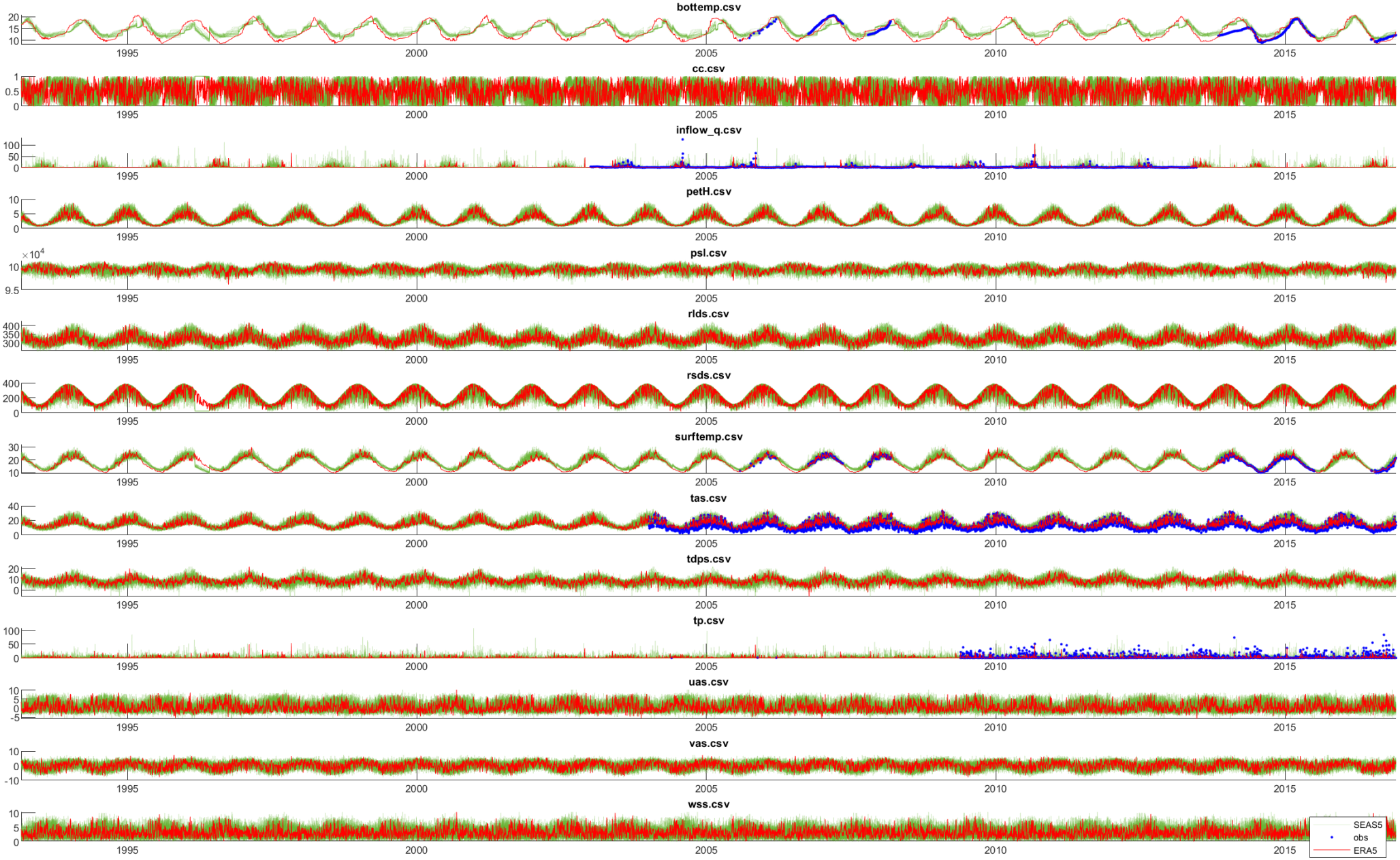

Fig S8: Mt Bold – Data coverage for hydrological and lake modelling, observations versus ERA5