Journal of Hydrology: Regional Studies 5 (2016) 226–243

Contents lists available at ScienceDirect

Journal of Hydrology: Regional Studies

journal homepage: www.elsevier.com/locate/ejrh

Water temperature forecasting for Spanish rivers by means of nonlinear mixed models

Yiannis Kamarianakis[a,*], Sergio Velasco Ayuso[b,c], Elena Cristóbal Rodríguez[d], Manuel Toro Velasco[c]

[a] School of Mathematical & Statistical Sciences, Arizona State University, Tempe, AZ 85287, USA
[b] School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA
[c] Center for Hydrographic Studies (CEDEX), Madrid 28005, Spain
[d] Center for Harbours and Coastal Studies (CEDEX), Madrid 28026, Spain

ARTICLE INFO

Article history:
Received 22 August 2015
Received in revised form 11 January 2016
Accepted 17 January 2016
Available online 13 February 2016

Keywords:
River temperatures
Streams
Predictive modeling
Thermal regimes
Spanish rivers

ABSTRACT

Study region: 43 rivers in Spain with measurement stations for air and water temperatures.
Study focus: River water temperatures influence aquatic ecosystem dynamics. This work aims to develop transferable river temperature forecasting models, which are not confined to sites with historical measurements of air and water temperatures. For that purpose, we estimate nonlinear mixed models (NLMM), which are based on site-specific time-series models and account for seasonality and S-shaped air-to-water temperature associations. A detailed evaluation of the short-term forecasting performance of both NLMM and site-specific models is undertaken. Measurements from 31 measurement sites were used to estimate model parameters whereas data from 12 additional sites were used solely for the evaluation of NLMM.
New hydrological insights for the region: Mixed models achieve levels of accuracy analogous to linear site-specific time-series regressions. Nonlinear site-specific models attain 1-day ahead forecasting accuracy close to 1 °C in terms of mean absolute error (MAE) and root mean square error (RMSE). Our results may facilitate adaptive management of freshwater resources in Spain in accordance with European water policy directives.

1. Introduction

In rivers, variations in water temperature influence processes across all levels of organization, from individual organisms to the ecosystem scale (Caissie, 2006). Temperature is associated with most physical and chemical properties of flowing waters and regulates interactions among the compartments that constitute a lotic ecosystem. For example, river water temperature impacts on dissolved oxygen concentration (Allan, 1995) and metabolic rates of various organisms (Brey, 2010). Furthermore it affects growth, emergence and survivorship of invertebrates, predator-prey encounter rates, interaction strengths (Rall et al., 2010), and the spatio-temporal patterns of invertebrate and fish assemblages (Gustafson, 2008).

Recent research has shown that some cold-water fishes are endangered by unsuitably warm temperatures in a way that further warming would result in net habitat loss (Isaak et al., 2010; Wenger et al., 2011). In addition, changes in thermal regimes present negative economic impacts in fisheries (Hague and Patterson, 2014). Understanding the dynamics

* Corresponding author.
E-mail address: yiannis76@asu.edu (Y. Kamarianakis).

http://dx.doi.org/10.1016/j.ejrh.2016.01.003

of water temperatures in rivers and their association with climatic conditions, topography and human activities, will enhance prediction of thermal shifts and will facilitate management of freshwater biodiversity (IPCC, 2013) in accordance with water policy directives such as the European Water Framework Directive (Directive 2000/60/EC of the European Parliament).

During the last 20 years, researchers have developed a wide variety of short-term forecasting models for stream water temperatures. Models fall in two categories: (i) deterministic (Caissie et al., 2007; Benyahya et al., 2010), and (ii) statistical (Caissie et al., 1998; Caissie et al., 2001; Benyahya et al., 2007). Deterministic models quantify energy fluxes between the river and its surroundings (atmospheric and streambed) and calculate water temperature variability over specific time scales using a heat budget approach (Caissie, 2006). Once calibrated for a given region, such models can be applied to different areas only if the 'new' physical characteristics are known. However, the large amount of data necessary, and the time and expense required to their development, hinders their wide use (Benyahya et al., 2007).

Statistical models are simple, require fewer input data and can be widely applied, given that air temperature sampling stations are commonly available (Caissie et al., 1998; Caissie, 2006). Time-series models achieve short-term forecasting accuracy comparable to deterministic models (i.e., errors within 1–2 °C for daily time steps; Caissie, 2006). Moreover, such models have been used to forecast water temperatures in rivers based on meteorological scenarios derived from Global Climate Models (Stefan and Sinokrot, 1993; Mohseni et al., 2003; Mantua et al., 2010; Jeong et al., 2013). A drawback of the statistical approaches, emphasized by Caissie (2006), is lack of transferability: multiple site forecasts from statistical models are carried out independently and are confined to locations with available historical data on air and water temperatures.

This work aims to address the aforementioned drawback by examining the generalization capability of a new model class in river temperature forecasting, namely nonlinear mixed models (NLMM), in regions with insufficient data for site-specific models. The adopted methodology focuses on short-term forecasting and is not site-specific. The dynamics of water temperatures in a set of measurement sites are described by a general model with parameters that may depend on environmental and geographical factors such as maximum or minimum annual temperature and elevation. Guillemette et al. (2009) and Daigle et al. (2010) also presented models that collectively analyze river temperatures from multiple measurement sites. The abovementioned works were based on different statistical tools (kriging and multivariate regressions, respectively) relative to the ones presented herein, and provided forecasts at the monthly level.

Mixed models have been applied in a wide variety of disciplines; they are useful in settings where repeated measurements are made on the same statistical units. In this article, NLMM were based on nonlinear site-specific models, similar to the ones proposed by Mohseni et al. (1998), Caissie et al. (2001), Larnier et al. (2010) and Hague and Patterson, (2014). Prior to NLMM development, the short-term forecasting performance of alternative site-specific models was evaluated and the dependence of their parameters on environmental and geographical factors was examined. This research focuses on the Iberian Peninsula, a region with a high biodiversity threatened by climate change (Thuiller et al., 2011). Our analyses used daily measurements from 43 temperature stations, each station located at a different river.

2. Methods

2.1. Study area

The Iberian Peninsula is located between the Atlantic Ocean and the Mediterranean Sea, and bridges the African and European continents (Fig. 1). This unique position, combined with a complex orography, explains its important climatic diversity (MIMAM, 2004). In general, the northern Iberian Peninsula shows mild winters and summers with annual rainfall between 800 and 1500 mm, average temperatures of 9 °C in winter and 18 °C in summer. The southern part has mild winters and hot summers (annual rainfall between 250 and 600 mm and average temperatures of 11 °C in winter and 23 °C in summer), and the central areas experience cold winters and hot summers (annual rainfall typically above 400 mm and average temperatures of 4 °C in winter and 24 °C in summer). The region can be roughly divided into humid (>750 mm/year), dry (∼350–750 mm/year), and arid areas (<350 mm/year). The humid area is found in the north part of the peninsula and in high-elevation ranges. The dry area is the largest, including the Guadalquivir and Ebro depressions, both Central Plateaus and most part of the Mediterranean coast. The arid zone is relatively small and constrained to the southeast.

Iberian rivers can be grouped in those flowing into the Atlantic Ocean and those flowing into the Mediterranean Sea. Five out of the main eight rivers of Spain flow directly into the Atlantic Ocean (Miño, Duero, Tajo, Guadiana and Guadalquivir) whereas three flow into the Mediterranean Sea (Ebro, Júcar and Segura). The orography of the territory determines the length and characteristics of the rivers, with Atlantic rivers being longer than Mediterranean rivers (MIMAM, 2004). The latter are typically torrential and highly irregular in terms of flow. Rivers flowing to the Cantabric Sea (north of the Iberian Peninsula) are short but carry large amounts of water because of strong and frequent rainfalls. Water from the Pyrenees feeds the Ebro River in the northeast, whereas water from the Baetic Range feeds the Guadalquivir River in the south part of Spain. In summary, climatic, geological, geomorphological, lithological and historical diversity provide with heterogeneity and recognizable particularities the hydrography of Spain, which can be considered one of the most varied in Europe (MIMAM, 2004).

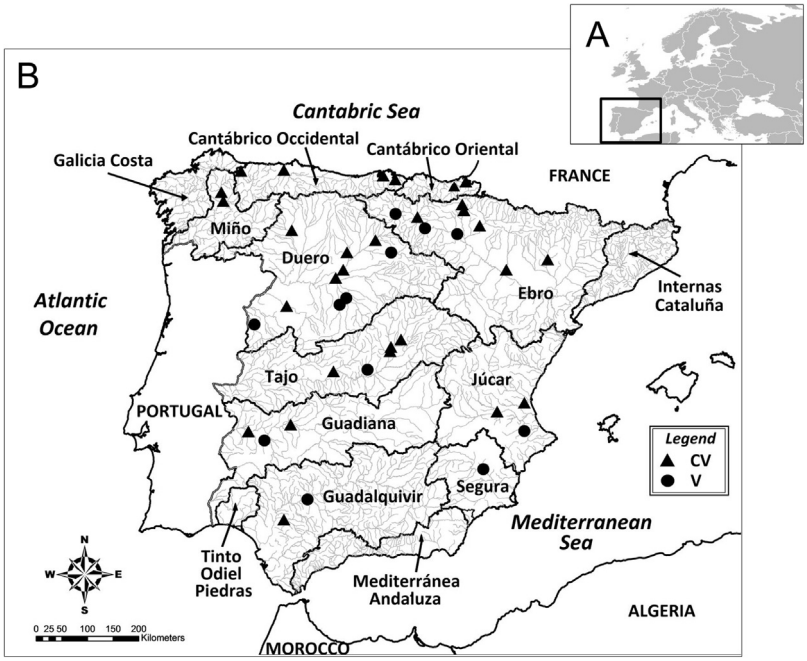

Fig. 1. (A) Location of the Iberian Peninsula in Europe. (B) SAICA measurement locations in the Iberian Peninsula. Triangles denote sites used both to calibrate and validate (CV) location-specific and mixed models and dots denote sites used solely to validate (V) the predictive power of mixed models.

2.2. Stream water and air temperature datasets

Water temperature time series (T_w) were obtained from 43 SAICA stations (Water Quality Automatic Information System, Spanish Environmental Department), whereas air temperature time series (T_a) were obtained from 43 AEMET stations (Spanish Meteorological Agency). Pairs of SAICA and AEMET stations were selected following a set of restrictive rules: (1) all SAICA stations had to be located at least 30 km downstream of large dams, large cities or fuel/nuclear power stations; (2) all AEMET stations had to be located less than 20 km of the corresponding SAICA stations (the average distance is 5.4 km with a standard deviation of 3.9 km) and (3) the altitudinal difference between pairs of stations had to be minimal (the average altitudinal difference is 35 m) with no mountain range separating them.

Our analysis was divided in three stages which used different sets of data. In the first stage forecasting models for 31 measurement sites, located in different rivers, were calibrated using daily data from three annual periods and validated on a 1-year period. Specifically, for each site, data from 2002 to 2008 were available; the 3 years with the most observations were used for model calibration. Annual periods with less than 300 daily measurements, which correspond to 82% of the annual data, were excluded to avoid estimation bias related to non-random patterns of missing data. Forecasting performance in each site was evaluated using data from the year with the 4th largest number (typically > 300) of daily measurements.

In stage 2, model parameters from stage 1 were associated with site-specific geographical and environmental factors. Finally, in the third stage, the dataset used for model estimation in stage 1 and the findings from stage 2, were used in the model building procedure of NLMM. Mixed models were validated on measurements collected from 12 rivers, different to the ones used in the first stage (the available data for these 12 sites were not sufficient to conduct analysis of the type performed in the first stage). This evaluation used data that correspond to the 1-year period with the largest number of measurements (typically > 300) per site.

Fig. 1 depicts locations of the 31 pairs of stations used for calibration and validation of site-specific models and NLMM (calibration/validation group, denoted by CV) as well as the 12 stations used solely to validate NLMM (validation group, denoted by V). Table 1 shows information on the European Water Framework Directive (WFD) ecological type of each water body (this explanatory variable is based on geographic, climatic, and other fundamental physical features), distances between pairs of water and air temperature measurement sites, and elevations of measurement sites. Table 1 also depicts which sites were used solely to validate the generalization ability of mixed models. Table 2 reports calibration and validation periods per site and summary statistics for air and water temperatures. Maximum water temperatures were greater than 20 °C in 41 measurement sites and exceeded 30 °C in 11 sites. In contrast with air temperatures, observed minimum water temperatures remained positive.

Table 1
Basic topographic characteristics of stream temperature measurement locations: river names, Water Framework Directive ecological types (WFD), River Basin name, distances (DIST) between measurement stations and elevations of stream (EL_W) and water (EL_A) temperature measurement stations. Group CV denotes locations for which both model calibration and validation were performed; measurements taken from sites in Group V were used solely for validation of NLMM.

GROUP	RIVER	WFD	BASIN	DIST [km]	EL_W [m]	EL_A [m]
CV	Narcea	128	Norte	3.82	79	85
CV	Asón	129	Norte	3.71	38	80
CV	Miño	128	Norte	0.68	406	440
CV	Bidasoa	129	Norte	3.86	15	35
CV	Oria	129	Norte	2.87	63	75
CV	Miera	132	Norte	3.77	14	34
CV	Eo	128	Norte	6.98	68	340
CV	Neira	128	Norte	3.63	354	400
CV	Pisuerga	115	Duero	5.42	698	700
CV	Duero	117	Duero	7.04	676	735
CV	Tormes	117	Duero	18.61	740	816
CV	Carrión	115	Duero	4.86	737	760
CV	Arlanzón	112	Duero	8.02	800	831
CV	Órbigo	115	Duero	7.42	810	844
CV	Manzanares	115	Tajo	1.94	536	530
CV	Henares	116	Tajo	1.29	599	610
CV	Tajo	117	Tajo	11.28	366	380
CV	Jarama	115	Tajo	7.08	553	582
CV	Zújar	117	Guadiana	6.12	257	302
CV	Guadajira	101	Guadiana	12.47	189	260
CV	Guadiana	101	Guadiana	12.47	188	260
CV	Guadaira	102	Guadalquivir	9.05	23	35
CV	Túria	118	Júcar	5.15	75	30
CV	Júcar	116	Júcar	2.46	414	400
CV	Arga	115	Ebro	0.83	394	387
CV	Ebro	115	Ebro	5.35	450	479
CV	Zadorra	115	Ebro	2.88	455	458
CV	Cinca	115	Ebro	0.36	249	243
CV	Aragón	115	Ebro	4.32	421	435
CV	Gállego	115	Ebro	1.02	220	225
CV	Arakil	126	Ebro	2.43	410	442
V	Arlanza	115	Duero	1.86	889	940
V	Agueda	115	Duero	2.12	655	680
V	Eresma	104	Duero	9.8	771	820
V	Adaja	104	Duero	7.3	789	820
V	Guadarrama	115	Tajo	9.32	475	500
V	Bonhaval	101	Guadiana	7	330	335
V	Guadalquivir	117	Guadalquivir	9.03	45	65
V	Segura	114	Segure	4.1	180	265
V	Albaida	109	Júcar	7.77	172	333
V	Ega	115	Ebro	0.58	304	306
V	Oca	112	Ebro	0.65	574	598
V	Najerilla	112	Ebro	5.32	437	437

2.3. Baseline models

2.3.1. Model C

Site-specific short-term forecasting models for water temperatures were based on two nonlinear associations. The first association focuses on annual temperature cycles and will be denoted as model C, since to our knowledge it was first adopted by Cluis (1972):

$$C: \hat{T}_w(t) = \alpha_1 + \alpha_2 \sin\left(\frac{2\pi}{365}(t+\tau)\right) \quad (1)$$

In (1) $\hat{T}_w(t)$ denotes predicted average daily water temperature on day t, whereas α_1, α_2, τ, represent unknown coefficients that should be estimated for each site. Site-specific subscripts on $\hat{T}_w(t)$ and the unknown coefficients are henceforth omitted for brevity. It is worth noting that estimation of annual cycles using (1) is parsimonious relative to models that are based on series of sinusoids and cosines for the long-term components (Ahmadi-Nedushan et al., 2007; Larnier et al., 2007; Hague and Patterson, 2014). These models were not adopted here as preliminary experiments showed that they would lead to overly complicated mixed models in the next stages with only small gains in forecasting accuracy. Furthermore the small number of parameters in (1) facilitates convergence of NLMM estimation algorithms.

Table 2
Calibration (C.; 3 years per location) and validation (V.; 1 year per location) periods. The calibration periods typically do not use data from consecutive years due to large numbers of missing measurements in some years. Columns 4–11 report minimum, median, mean, maximum of observed water and air temperatures respectively, observed in the model-calibration (evaluation) period for Group CV (V) sites.

River	C.	V.	Min_w [°C]	Med_w [°C]	Mean_w [°C]	Max_w [°C]	Min_a [°C]	Med_a [°C]	Mean_a [°C]	Max_a [°C]
Narcea	03,07,08	02	6.1	13.4	14.1	15.7	−4.6	14.2	14	37.6
Asón	06,07,08	02	4.7	12.1	13.2	26.8	−3	14.2	14	37.3
Miño	02,04,05	06	1.6	12.8	13.8	30	−9	12	12.2	35
Bidasoa	02,04,05	06	4.9	14.2	14.4	23.8	−5.4	15	14.6	37.2
Oria	00,07,08	06	4.2	14	14.6	27.5	−6	14.5	14.1	37
Miera	06,07,08	02	5.5	14.2	14.9	28.2	−4	13.8	13.5	36
Eo	02,04,08	05	5	11.9	12.8	23.2	−6.5	13.5	13.3	34
Neira	02,04,07	05	1.9	13.2	13.6	24.8	−8	13.5	13	37
Pisuerga	05,06,08	07	2.6	15	14.7	27.7	−11.5	12.9	12.8	37.5
Duero	06,07,08	05	4.1	15.3	15.1	28	−6.8	12.4	12.6	38.2
Tormes	06,07,08	04	3.5	16.9	16.6	30.8	−10	11	11.3	38
Carrión	06,07,08	05	1.9	14.1	13.4	24.9	−8	12	12.2	36
Arlanzón	06,07,08	05	2.6	12.6	13	30	−9	10.5	11	37.5
Órbigo	06,07,08	05	1	12.4	11.8	23.2	−8	10.5	10.7	34
Manzanares	05.06,08	07	10.8	20.9	21.2	42.5	−12	15.2	15.6	40.2
Henares	05,06,07	04	2.1	16.1	15.4	26.8	−12	13	13.1	40
Tajo	05,06,07	08	4	19.5	19.6	39.7	−8	16	16.3	40
Jarama	05,06,07	09	4.7	17.6	17.2	31.3	−9.1	14.6	15.1	40.7
Zújar	06,07,08	04	6.7	19.4	18.8	32.4	−5	18	17.9	43
Guadajira	06,07,08	04	5.8	18.2	17.8	30.3	1	16	16.2	41
Guadiana	06,07,08	04	6.7	19.6	19.1	32.3	1	16	16.3	41
Guadaira	05,07,08	06	1.4	19.8	19.8	40	−6	18.5	18.8	42
Túria	06,07,08	03	1.4	17.6	17.8	29.2	−3	18	18.2	39
Júcar	06,07,08	04	7.3	18.4	17.6	26.9	−8.1	16.1	16.3	41.7
Arga	05,07,08	06	1	13.1	13.9	27	−13.5	13	12.8	38
Ebro	06,07,08	05	6.8	16.3	16.5	28.6	−6	13.8	13.8	38
Zadorra	06,07,08	05	4.3	13.7	13.8	25.4	−7	13	12.9	36.5
Cinca	06,07,08	04	5.5	14.2	14.7	26.8	−6	14.5	14.7	37
Aragón	06,07,08	04	3.7	13.4	13.1	23.1	−6.5	14	14	37.5
Gállego	05,06,08	04	1.8	15.7	15.33	28.9	−10.7	14.9	14.6	40.4
Arakil	05,06,08	04	2.5	12.7	13.3	27.4	−8	13.8	13.1	37.5
Arlanza		08	1	10.8	11.5	24.1	−8	11.5	12.4	37
Agueda		05	2	13.1	13.1	28.9	−7	9.5	9.5	34
Eresma		05	1	11.8	12.1	26.3	−11	12	11.9	38.7
Adaja		08	3	11.7	12.5	25.5	−5.1	11.4	12.1	38
Guadarrama		05	1.6	17.3	16.9	31.8	−8	14.8	15.1	44
Bonhaval		04	6.6	15.1	16.4	27.3	−3	14	16.4	41.6
Guadalquivir		05	5.1	17.8	16.8	26.2	−6.5	19.5	18.9	45
Segura		07	7.3	17.1	17.4	33.4	−2.4	16.4	17.5	41.1
Albaida		07	8.8	17.6	18.5	31.5	1.5	18	18.3	39
Ega		02	5.2	13.1	13.7	24	−3	14.8	14.9	38.5
Oca		02	3.9	10.8	10.9	18.5	−2	12.5	12.9	38
Najerilla		07	3.5	12.3	12.6	21.7	−6	12.5	13	36

2.3.2. Model M

The second baseline model is based on the logistic function and implies an S-shaped association between air and water temperatures:

$$M : \hat{T}_w(t) = \mu_1 + \frac{\mu_2 - \mu_1}{e^{(\mu_3 - \mu_4 T_a(t))}} \tag{2}$$

In (2), $T_a(t)$ denotes air temperatures observed on day t, μ_1 is the 'floor' that corresponds to the lowest predicted stream temperature levels, μ_2 is the 'ceiling' (maximum predicted stream temperature) of the estimated association at a site, $\mu_4$is a measure of the steepest slope of the logistic function and the ratio μ_3/μ_4represents air temperature at the inflection point of the S-shaped profile. The logistic model in (2) has been proposed by Mohseni et al. (1998) and will be denoted by M.

2.3.3. Model L

A third model which is evaluated against the sinusoidal and the logistic, implies a linear air-to-water temperature association and hereafter will be designated by L:

$$L : \hat{T}_w(t) = \kappa_0 + \kappa_1 T_a(t) \tag{3}$$

Such linear associations have been estimated in several research works; a recent example is presented in Arismendi et al. (2014). L was not used as the basis of more advanced site-specific short-term forecasting models as an initial examination,

presented in Section 3.1, revealed inferior performance on the testing subset relative to the nonlinear specifications presented above.

Unknown coefficients in (1)–(2), were estimated using (a) the Gauss-Newton nonlinear least squares algorithm and (b) the interior point algorithm for nonlinear median regression (Koenker and Park, 1994). The latter algorithm minimizes the sum of absolute deviations of the residuals and may lead to superior forecasting performance as estimation based on least absolute deviations is robust to outliers and skewed response distributions (Dielman, 1986). Coefficients of the linear model in (3) were also estimated using least squares (LS) and least absolute deviations (LAD) algorithms.

2.4. Site-specific short-term forecasting models

A family of short-term water temperature forecasting models, presented in Caissie et al., 1998, can be based on (a) deviations of water temperatures from their long-term annual components presented in (1), and (b) deviations of air temperatures T_a from their corresponding long-term components presented below:

$$\hat{T}_w(t) = \alpha_{1,\alpha} + \alpha_{2,\alpha}\sin\left(\frac{2\pi}{365}(\mathrm{t}+\tau_\alpha)\right) \tag{4}$$

Let $R_{w,1}$ denote observed deviations of water and R_a observed deviations of air temperatures respectively. Two time-series regression models dubbed C_a and C_b, based on time-lags of $R_{w,1}$ and R_a, can be formulated as:

$$C_a : \mathcal{R}_{w,1}(t) = \theta_0 + \sum_{\mathrm{j}=1}^{p}\theta_{1,i}R_a(t-j) \tag{5}$$

$$C_b : \mathcal{R}_{w,1}(\mathrm{t}) = \theta_0 + \sum_{\mathrm{j}=1}^{p}\theta_{1,j}R_a(t-j) + + \sum_{\mathrm{i}=1}^{q}\theta_{2,i}R_{w,1}(t-i) \tag{6}$$

with i, j denoting time lags of deviations of water and air temperatures, respectively. The above models produce 1-day ahead forecasts: extensions for multiple-day-ahead forecasting are straightforward but will not be pursued in this article.

A second family of short-term forecasting models, which to our knowledge is examined here for the first time, can be based on deviations of observed water temperatures from their site-specific air-to-water association profiles (model M). Let $R_{w,2}$ denote these deviations; three time series models, dubbed respectively M_a, M_b, M_c can be formulated as follows:

$$M_a : \mathcal{R}_{w,2}(t) = \theta_0 + \sum_{\mathrm{i}=1}^{p}\theta_{1,i}R_{w,2}(t-i) \tag{7}$$

$$M_b : \mathcal{R}_{w,2}(t) = \theta_0 + \sum_{\mathrm{j}=1}^{q}\theta_{2,j}R_a(t-j) \tag{8}$$

$$M_c : \mathcal{R}_{w,2}(t) = \theta_0 + \sum_{\mathrm{i}=1}^{p}\theta_{1,i}R_{w,2}(t-1) + \sum_{\mathrm{j}=1}^{q}\theta_{2,j}R_a(t-j) \tag{9}$$

In (7)–(9) R_a represents observed deviations of air temperatures from their long-term annual components, as in (5) and (6).

Unknown coefficients in (5)–(9) were estimated using both LS and LAD algorithms. Autoregressive orders p and q where chosen based on a forward stepwise procedure and Akaike's Information Criterion (Hyndman and Khandakar, 2008). The adequacy of linear autoregressive models was investigated by applying a battery of specification tests, including the Ljung-Box and Breusch-Godfrey tests of residual autocorrelation (Ljung and Box, 1978; Godfrey 1978) the Keenan and Tsay tests of remaining nonlinearity (Keenan 1985; Tsay, 1986) and the Breusch-Pagan test of heteroscedasticity (Breusch and Pagan, 1979). All calculations were performed in R (R Development Core Team, 2014).

2.5. Nonlinear mixed models

The time-series models presented in the previous sections treat each site independently. Multilevel, or mixed effects models can be seen as generalizations of site-specific models which summarize water temperature dynamics when measurements are collected from a group of sites. Mixed models contain fixed and random effects: fixed effects can be viewed as 'average' model parameters whereas random effects represent site-specific deviations from 'average' dynamics. Nonlinear mixed effects models (NLMM) can be used for site-specific inference and forecasting when estimates at both levels (fixed and random effects) are taken into account. NLMM can be used to forecast water temperatures at any site, based solely on the coefficients that correspond to fixed effects; this includes sites for which site-specific model calibration is infeasible due to insufficient data.

The NLMM dubbed C^*, based on C in (1), is formulated as:

$$\begin{aligned} C^*: T_{w,s}(t) &= \beta_1 + b_{1,s} + \left(\beta_2 + b_{2,s}\right) \sin\left(\frac{2\pi}{365}\left(t + \left(\beta_3 + b_{3,s}\right)\right)\right) + \varepsilon_s(t) \\ b_{j,s} &\sim N\left(0, \psi_j^2\right), \quad Cov\left(b_{j,s}, b_{j',s}\right) = \psi_{jj'} \\ \varepsilon_s(t) &\sim N\left(0, \sigma^2\lambda_s\right), \quad Cov\left(\varepsilon_s(t), \varepsilon_s(t-k)\right) = \sigma^2\lambda_{sk} \end{aligned} \tag{10}$$

The fixed effects coefficients β_1, β_2 and $\beta_3$represent the mean values of the parameters of C across all measurement sites. Site-specific deviations are represented by the random effects $b_{j,s}, j = 1, 2, 3$with s denoting measurement site. Random effects are assumed to follow a multivariate normal distribution with variances ψ_j^2 and covariances $\psi_{jj'}$; Ψwill denote the resulting covariance matrix. Random effects corresponding to different measurement sites are assumed to be independent. The site-specific error terms ε_s, are assumed normally distributed and independent of the random effects. The covariance structure of the error terms aims to capture serial correlation: the λ_{sk} (k denotes time lags) are specified to correspond to an autoregressive moving average specification ARMA(p,q).

C_a^*, the NLMM analogue of C_a is similar to (10) with additional terms

$$\left(\beta_4 + b_{4,s}\right) R_a(t-1) + \ldots + \left(\beta_k + b_{k,s}\right) R_a(t-k), \tag{11}$$

with k representing time lags of air temperature deviations from their seasonal profiles. Similarly, M^* is formulated as

$$\begin{aligned} M^*: \widehat{T_{w,s}}(t) &= \delta_{1,s} + d_{1,s} + \frac{\delta_{2,s} + d_{2,s} - \delta_{1,s} - d_{1,s}}{\exp\left\{\delta_{3,s} + d_{3,s} + \left(\delta_{4,s} + d_{3,s}\right) T_a(t)\right\}} + \varepsilon_s(t) \\ d_{j,s} &\sim N\left(0, \psi_j^2\right), \quad Cov\left(d_{j,s}, d_{j',s}\right) = \psi_{jj'} \\ \varepsilon_s(t) &\sim N\left(0, \sigma^2\lambda_s\right), \quad Cov\left(\varepsilon_s(t), \varepsilon_s(t-k)\right) = \sigma^2\lambda_{sk} \end{aligned} \tag{12}$$

with δ_i, $d_i i = 1, ..4$,denoting fixed and random effects respectively and the same assumptions as in (10) with regard to the distributions of the residuals and the random effects.

The initial step in the model building procedure of mixed effects models is deciding which of the coefficients need random effects to account for their between-site variation and which can be treated as purely fixed effects. The procedure applied herein started with a model with random effects for all parameters and an un-constrained covariance structure. Simplified versions of the general model with fewer random effects and block-diagonal covariance structures were evaluated using likelihood ratio tests [these tests compare nested models estimated by maximum likelihood; further details are presented in Pinheiro and Bates (2009) and the Bayesian Information Criterion (BIC; Zuur et al., 2009). With regard to the error terms $\varepsilon_s(t)$, two alternatives were examined: the first targeted well-specified models and estimated the autoregressive and moving average order of the ARMA specification using a backward stepwise procedure with maximum values 5 and 3 respectively. The second model building strategy did not account for serial correlation in residuals, given that water temperatures are assumed unknown in the evaluation: forecasts are based only on short-term information on air temperatures (this renders our models transferable to location without measurements on water temperatures) and residual autoregressive dynamics cannot be used to improve forecast accuracy.

In the next stage of the model building process, part of the variability on the random effects in (10)–(12), was explained by site-specific environmental and geographical characteristics. Specifically, a forward stepwise procedure was implemented to evaluate whether terms $\beta_j + b_{j,s}, j = 1, \ldots, 3$ in (10) should be augmented with information on site-specific covariates: $\beta_j + b_{j,s} + \phi_j^1 X_{1,s} + \ldots + \phi_j^m X_{m,s}$; the same procedure was followed for the random effects in (11) and (12). The covariates used were average, median, minimum and maximum annual air temperatures, elevation of the water temperature measurement site and a dummy variable that indicates whether a site belongs to the most frequently observed WFD type. The effects of the remaining WFD types were not examined due to the small number of rivers per type. The forward stepwise procedure was based on likelihood ratio tests and BIC, and was guided from exploratory scatterplots of random effects on each $X_{i,s}$. After this step, random effects were re-evaluated and a final specification was derived.

The forecasting performance of three NLMM variants of C, C_a, M, M_b, which produce forecasts without using short-term information on water temperatures, was evaluated on 12 measurement sites. These sites were not part of the data used in the model building process. Mixed models analogous to the ones shown in (10)–(12), without autoregressive dynamics for the residuals, were compared to models which use environmental and geographical covariates to explain part of the variability of the random effects and to models which also take into account residual autocorrelation. The assumptions of variance homogeneity of site-specific residuals, normality of site-specific residuals, normality of random effects and the correlations of residuals and random effects were examined using graphical tools and statistical information criteria (Pinheiro and Bates, 2009; Zuur et al., 2009).

Table 3
Average (across measurement locations) MAE, RMSE, Type I and Type II error rates for the predictive models in the evaluation period. Performance of models estimated using LAD is shown in parentheses.

	C	C_a	C_b	M	M_a	M_b	M_c	L
$\overline{\text{MAE}}$ [°C]	1.62(1.63)	1.15(1.15)	0.41(0.41)	1.70(1.69)	1.05(1.06)	1.62(1.61)	1.00(1.01)	1.74(1.74)
$\overline{\text{RMSE}}$ [°C]	2.03(2.05)	1.46(1.46)	0.57(0.56)	2.14(2.14)	1.38(1.41)	2.03(2.03)	1.32(1.35)	2.18(2.18)
$\bar{E}_1$	0.46(0.45)	0.33(0.33)	0.11(0.11)	0.41(0.39)	0.21(0.21)	0.39(0.37)	0.17(0.17)	0.41(0.40)
$\bar{E}_2$	0.05(0.05)	0.03(0.04)	0.01(0.01)	0.04(0.03)	0.03(0.03)	0.03(0.04)	0.03(0.03)	0.03(0.03)

2.6. Evaluation metrics

Water temperature forecasts from NLMM and site-specific models were compared with historical observations of daily water temperatures in the evaluation period using five performance criteria: (1) mean error (ME) which is a measure of bias; (2) mean absolute error (MAE); (3) root mean square error (RMSE); (4) Type I error rate and (5) Type II error rate, both relative to a threshold of 20 °C, chosen to represent the biological preferences of typical cold-stenotherm organisms (Lessard and Hayes, 2003). Type I error occurs when the observed water temperature exceeds the threshold but the model fails to predict it. Conversely, Type II error occurs when observed water temperature is below the designated threshold but the model erroneously predicts that it is above.

3. Results and discussion

3.1. Baseline models: predictive performance

Fig. 2 depicts the distributions of the five performance metrics for baseline models C, M and L, estimated using LS and LAD. The performance of median regression models in the evaluation period was found similar to the one achieved by least-squares regression, which suggests the absence of outlying measurements in the calibration period (Table 3). Indeed, LAD-estimated curves were almost identical to the ones estimated by conventional least squares with minor deviations for model M (Fig. 3).

Interestingly, predictive models based solely on seasonal cycles (model C) performed better in the evaluation period in terms of MAE and RMSE (Fig. 2b,c) than site-specific models based on air-to-water association profiles (M), despite the fact that the former did not use information on air temperatures. In accordance with prior expectations, linear air-to-water relationships (L) performed worse than both C and M; furthermore, the same performance ranking across models was observed when median MAE and median RMSE were used as summary measures of accuracy across sites. It is worth noting that C performed better than M in terms of both MAE and RMSE, in 19 (out of 31) measurement locations and L performed better than M in 7 locations. Moreover, C was the best performing model in 19 locations, M in 7 and L in 5.

Baseline models frequently underpredicted water temperatures above the 20 °C threshold: average (across sites) Type I error rates for M and L were close to 41% (Table 3) whereas model C performed even worse. On the other hand, C dominated based on median Type I error rate across sites (15% versus 18% and 17% for M and L respectively), which suggests that in some measurement sites C performed much worse than M and L (Fig. 2d). L was the best performing model in terms of both mean and median Type II error rate; false alarms appear to be a less significant problem for baseline models as Type II error rates (Fig. 2e) were substantially lower than Type I rates.

3.2. Short-term forecasting models

Fig. 4 depicts distributions of the five performance metrics across measurement sites for models C_a, C_b, M_a, M_b and M_c, estimated using LS; averages of MAE, RMSE, Type I and Type II error are shown in Table 3. C_b outperformed all alternative specifications across all sites in terms of all metrics, with average MAE (Fig. 4b) and RMSE (Fig. 4c) well below 1 °C. Type I error rates (Fig. 4d) improved dramatically relative to the baseline models and were close to 10% for C_b. M_c was the second best model with average RMSE across sites below 1.5 °C and average type I error rate below 20%. It is noteworthy that M_b, the model based on air-to-water association profiles and time lags of deviations of air temperatures from their annual components, performed poorly as its performance metrics were close to the ones of the baseline model C (for instance, average RMSE across sites was slightly above 2 °C).

The inferior performance of C_a and M_b, which did not use short-term information on water temperatures to produce 1-day ahead forecasts, is also manifested in Table 4 via the results of the Ljung-Box and Breusch-Godfrey hypothesis tests: these two models clearly failed to capture the autocorrelation of the $R_{w,1}$ (t) and $R_{w,2}$ (t) series. The two tests on remaining non-linearity (Table 4) suggest that in the vast majority of measurement sites, there was no evidence of remaining non-linearity in the residuals of C_b, M_a and M_c. Based on this finding, one should not expect improved forecasting performance by estimating nonlinear models (such as regime switching autoregressive specifications) to describe the dynamics of $R_{w,1}$ (t) and $R_{w,2}$ (t). Finally, the Breusch-Pagan test rejected the null hypothesis of homoscedastic residuals in about 50% of measurement sites for the best performing models C_b and M_c.

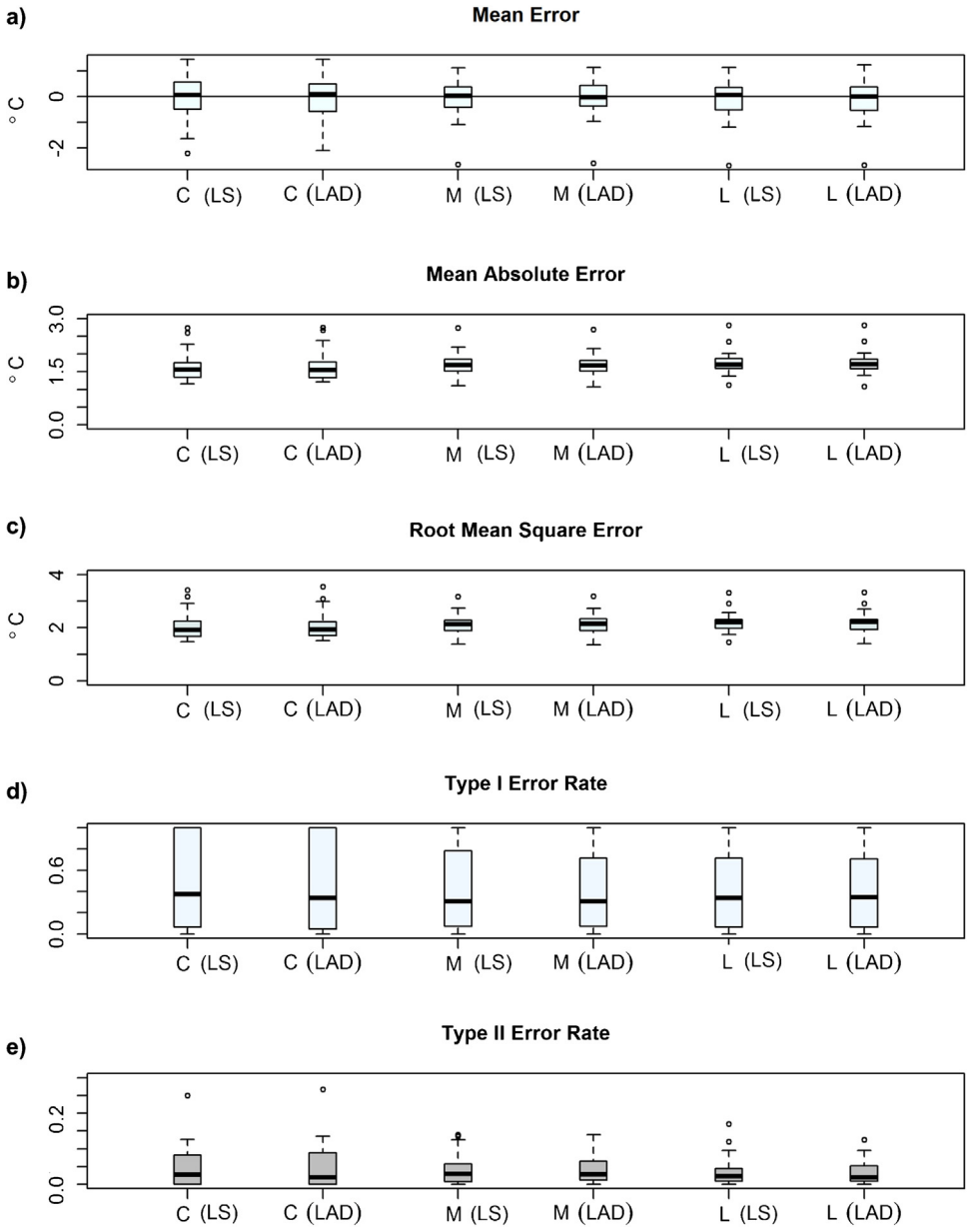

Fig. 2. Distributions of (a) mean error (bias); (b) MAE; (c) RMSE; (d) Type I and (e) Type II error rate across measurement sites for models *C*, *M* and *L* estimated by LS and LAD.

Table 4
Number of sites for which the null hypothesis of the specification tests is rejected at the 0.01 significance level. Tests were applied to model residuals in the calibration period. Rows correspond respectively to the Ljung-Box and Breusch-Godfrey tests for residual autocorrelation, Keenan's and Tsay's tests for remaining nonlinearity and the Breusch-Pagan test for heteroscedasticity.

	C_a	C_b	M_a	M_b	M_c
L-B	31	1	2	31	0
B-G	31	6	4	31	2
K	3	1	1	3	1
T	10	4	5	12	4
B-P	10	17	5	8	12

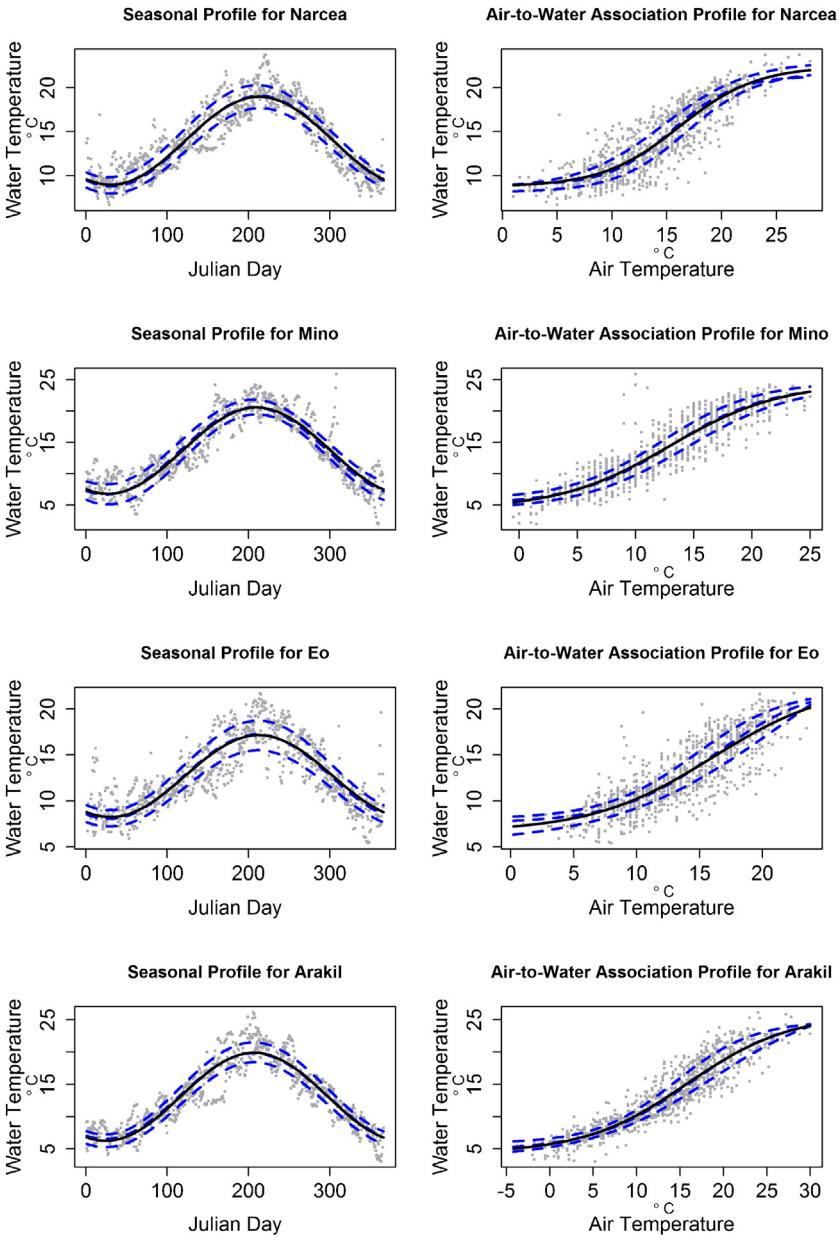

Fig. 3. Seasonal (model *C*, left) and air-to-water association profiles (model *M*, right) for four selected measurement locations. Data from the calibration period (3 years) are shown in grey. Solid curves correspond to LS estimates and middle dashed curves correspond to LAD. Outer dashed curves predict the 0.25 and 0.75 quantiles of water temperatures.

Our results confirmed previous research findings: site-specific statistical models can forecast river water temperatures with RMSEs close to 1 °C for a 1-day ahead horizon (Larnier et al., 2010; Hague and Patterson, 2014). In accordance with previous works, it was shown that models that exploit short-term information on water temperatures outperform specifications based solely on short-term dynamics of air temperatures. Variants of the nonlinear modeling approach proposed by Mohseni et al. (1998) were not found superior to the widely applied specifications based on seasonal decompositions. However, logistic models could be useful in forecast combination schemes, which improve the performance achieved by any single model (Pappenberger et al., 2015).

3.3. Site-specific models: parameter variability

Table 5 presents summary statistics for the parameters of *C*, C_a and *M*, which, given the estimated annual cycles, used solely short-term information on air temperatures to produce forecasts. Results for M_b and *L* are not shown, since based

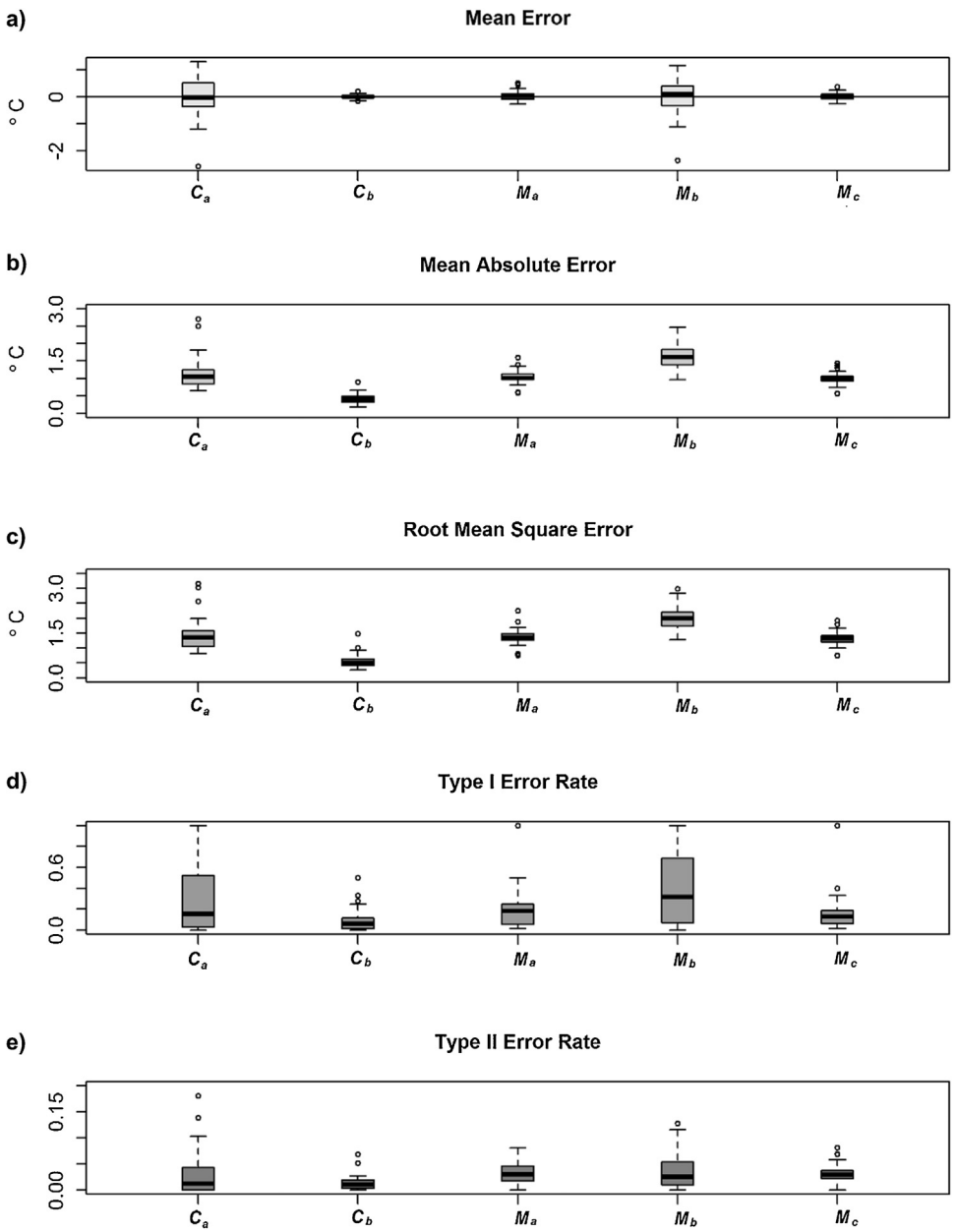

Fig. 4. Distributions of (a) mean error (bias); (b) MAE; (c) RMSE; (d) Type I and (e) Type II error rate across measurement sites for models C_a, C_b, M_a, M_b and M_c estimated by LS.

Table 5
Summary statistics of estimated (via least squares) parameters for models C, C_a and M.

	$\alpha 1$	$\alpha 2$	τ	θ_0	θ_1	θ_2	θ_3	μ_1	μ_2	μ_3	μ_4
Mean	15.40	−7.07	−296.2	0.02	0.30	0.08	0.11	5.65	25.70	2.84	0.20
Median	15.58	−7.06	−295.5	0.02	0.29	0.08	0.10	5.56	25.96	2.78	0.20
IQR	3.74	2.64	7.12	0.07	0.07	0.03	0.07	3.24	4.51	0.92	0.04
Max.	26.93	−4.46	−285.5	0.10	0.41	0.18	0.27	14.38	32.09	4.58	0.31
Min.	11.82	−9.34	−306.9	−0.10	0.10	−0.01	0.04	0.93	18.57	1.49	0.13

Table 6
Parameter estimates, residual sum of squares (RSS) and coefficient of determination (R^2) for linear models that aim to predict estimated (via least squares) coefficients in models C, C_a and M. The set of explanatory variables includes altitude of measurement site, a dummy variable that indicates sites that belong to the most frequently observed WFD type (115), minimum, maximum, mean, median air temperature in the calibration period and multiplicative interactions of all variables. All parameters are significant at the 0.05 level; empty cells correspond to variables that were not included in the final model produced from the stepwise linear model building procedure (based on BIC).

	α_1	α_2	τ	θ_0	θ_1	θ_2	θ_3	μ_1	μ_2	μ_3	μ_4
Intercept	−14.95	8.52	−354	0.02	4.05	0.07	0.09	5.59	−12	0.46	7.85
Min(T_a)						−0.02		1.61			
Med(T_a)	0.88										0.35
Mean(T_a)					−0.21						
Max(T_a)	0.55	−0.38	1.42		−0.09				0.99	−0.01	−0.26
Min(T_a): mean(T_a)						0.001		−0.12			
Max(T_a): mean(T_a)					0.01						
EL_W/1000	4.28	−2.43	8.19		−0.11		0.07				
RSS	1.25	0.98	4.28	0.05	0.06	0.04	0.05	2.11	2.25	0.04	0.58
R^2	0.76	0.53	0.43		0.29	0.15	0.12	0.39	0.51	0.16	0.44

on the results of the previous section, L was inferior to M whereas for M_b, the gains in predictive performance relative to M were not large enough to justify further investigation of its generalization potential. Mohseni et al. (1998) applied M on weekly data from 573 streams in the United States. The average of μ_1, the 'floors' of the air-to-water temperature association profiles across the 31 measurement locations of the calibration data, was about 5 °C higher than the value of the corresponding parameter reported in their paper. The average value of the ratioμ_3/μ_4, the air temperatures at the inflection points of the air-to-water association profiles, was about 1 °C higher in our sample. On the other hand, the averages of μ_2, the 'ceilings' of the profiles and μ_4, the steepest slopes of the sigmoids, were very close to the ones reported in Mohseni et al. (1998).

Associations of site-specific model parameters with geographical and environmental variables are shown in Table 6, which reports results from parameter-specific, forward-stepwise linear model building procedures. It can be observed that the proportion of parameter variability for C and M explained by altitude and summary statistics of air temperatures observed at nearby meteorological stations, ranged from 16% to 76%. Results were in accordance with intuition: variables included in the predictive model for τ were maximum air temperature and altitude, whereas information on maximum air temperature explained approximately 50% of the variability of the 'ceilings' in M. Interestingly, estimated parameters did not differ significantly for sites that belonged to the most frequently observed WFD type; the corresponding dummy variable was not found statistically significant in any parameter-specific model.

3.4. Mixed effects models: estimation

Eqs. (13)–(15) present, respectively, C^*, C^{**}, C^{***}: three NLMM based on C. In (13) one observes that estimated fixed effects were identical to the reported means in Table 4. Moreover, the model building procedure resulted in significant random effects with an unrestricted covariance structure for the three coefficients that characterize annual cycles in water temperatures.

$$C^*: \hat{T}_{w,s}(t) = 15.40 + b_{1,s} + \left(b_{2,s} - 7.07\right)\sin\left(\frac{2\pi}{365}\left(t + \left(b_{3,s} - 296.22\right)\right)\right) + \varepsilon_s(t)$$

$$\Psi_{1,1} = \begin{bmatrix} 2.37 & -0.48 & 0.45 \\ -0.48 & 1.35 & -0.61 \\ 0.45 & -0.61 & 5.35 \end{bmatrix} \quad \sigma = 1.87 \tag{13}$$

The off-diagonal elements of the Psi matrices in (13)–(20) correspond to estimated correlations instead of covariances. Correlations are more informative for model-building purposes, as strongly correlated random effects indicate over-parameterized models. Diagnostic checks were conducted on the normality assumption of random effects. Also the variance homogeneity and normality of site-specific residuals were assessed.

In C^{**}, shown in (14), part of the variability of the random effects $b_{1,s}$ and $b_{3,s}$ was explained by site-specific average and maximum annual air temperature denoted respectively $X_{1,s}$ and $X_{2,s}$)in the calibration period. This reduced substantially the variability of random effects as can be observed by a comparison of the diagonal elements of the covariance matrices $\Psi_{1,1}$ and $\Psi_{1,2}$. Hence, one should expect improved generalization ability relative to C^*.

$$C^{**}: \hat{T}_{w,s}(t) = 1.1X_{1,s} + b_{1,s} + \left(b_{2,s} - 7.04\right)\sin\left(\frac{2\pi}{365}\left(t + \left(0.72X_{2,s} + b_{3,s} - 323.71\right)\right)\right) + \varepsilon_s(t)$$

$$\Psi_{1,2} = \begin{bmatrix} 1.53 & -0.41 & 0.51 \\ -0.41 & 1.35 & -0.51 \\ 0.51 & -0.51 & 4.68 \end{bmatrix} \quad \sigma = 1.87 \tag{14}$$

C^{***}, shown in (15), accounted for autoregressive dynamics of the residuals:

$$C^{***}: \hat{T}_{w,s}(t) = 1.1X_{1,s} + b_{1,s} + \left(b_{2,s} - 7.14\right)\sin\left(\frac{2\pi}{365}\left(t + \left(b_{3,s} - 296.09\right)\right)\right) + \varepsilon_s(t)$$

$$\Psi_{1,3} = \begin{bmatrix} 1.57 & -0.31 & 0.44 \\ -0.31 & 1.56 & -0.56 \\ 0.44 & -0.56 & 4.64 \end{bmatrix} \quad u_s(t) \sim N\left(0, \sigma_2^2\right)$$

$$\varepsilon_s(t) = 1.16\varepsilon_s(t-1) - 0.39\varepsilon_s(t-2) + 0.12\varepsilon_s(t-3) + u_s(t), \quad \sigma_2 = 1.80 \tag{15}$$

One may observe differences in the estimated fixed-effects relative to C^{**}; these differences are expected to influence the forecasting evaluation that will follow, which was based solely on those parameters. Although C^{***} accounted for autocorrelated residuals (therefore it was correctly specified whereas C^{**} was not), the additional terms for residual dynamics were not used in the validation procedure; their application required knowledge of water temperatures, which were assumed unknown as these models should be transferable to sites with no measurements on water temperatures.

Eqs. (16)–(18) present, respectively, C_a^*, C_a^{**}, C_a^{***}; NLMM based on C_a. In (16) it can be seen that only the first lag of deviations of air temperatures from their seasonal profiles needed a random effect to account for site-specific variability; the coefficients of the remaining lags did not vary significantly across measurement sites:

$$C_a^*: \widehat{T_{w,s}}(t) = 15.42 + b_{1,s} + \left(b_{2,s} - 7.09\right)\sin\left(\frac{2\pi}{365}\left(t + \left(b_{3,s} - 296.21\right)\right)\right) + 0.28R_s^a(t-1)$$
$$+ b_{4,s} + 0.08R_s^a(t-2) + 0.04R_s^a(t-3) + 0.09R_s^a(t-4) + \varepsilon_s(t)$$

$$\Psi_{2,1} = \begin{bmatrix} 2.36 & -0.49 & 0.44 & -0.18 \\ -0.49 & 1.36 & -0.63 & -0.03 \\ 0.44 & -0.63 & 5.35 & 0.29 \\ -0.18 & -0.03 & 0.29 & 0.06 \end{bmatrix} \sigma = 1.33 \tag{16}$$

The standard deviation of the residuals, σ, decreased substantially relative to (13)–(15), which did not include information on air temperatures.

In C_a^{**}, part of the variability of the site-specific random effects $b_{1,s}$ and $b_{3,s}$ was explained by site-specific average and maximum annual air temperatures:

$$C_a^{**}: \widehat{T_{w,s}}(t) = 1.08X_{1,s} + b_{1,s} + \left(b_{2,s} - 7.07\right)\sin\left(\frac{2\pi}{365}\left(t + \left(0.71X_{2,s} + b_{3,s} - 323.43\right)\right)\right) + 0.28R_s^a(t-1)$$
$$+ 0.08R_s^a(t-2) + 0.04R_s^a(t-3) + 0.09R_s^a(t-4) + \varepsilon_s(t)$$

$$\Psi_{2,2} = \begin{bmatrix} 1.52 & -0.41 & 0.51 \\ -0.41 & 1.36 & -0.52 \\ 0.51 & -0.52 & 4.73 \end{bmatrix}, \sigma = 1.33 \tag{17}$$

This resulted in a significant reduction of the variability of the random effects, as can be verified by comparing the diagonal elements of $\Psi_{2,1}$ and $\Psi_{2,2}$, and is expected to lead to improved forecasting performance.

C_a^{***}, shown in (18), accounted for residual autocorrelation. The fixed-effects part of the model, which will be used to validate the generalization ability of estimated mixed models to sites for which model estimation did not take place, differed relative to C_a^{**} as it did not include information on maximum annual air temperatures:

$$C_a^{***}: \widehat{T_{w,s}}(t) = 1.09X_{1,s} + b_{1,s} + \left(b_{2,s} - 7.11\right)\sin\left(\frac{2\pi}{365}\left(t + \left(b_{3,s} - 295.95\right)\right)\right) + 0.12R_s^a(t-1)$$
$$+ 0.07R_s^a(t-2) + 0.02R_s^a(t-3) + 0.01R_s^a(t-4) + \varepsilon_s(t)$$

$$\Psi_{2,3} = \begin{bmatrix} 1.57 & -0.30 & 0.44 \\ -0.30 & 1.52 & -0.58 \\ 0.44 & -0.58 & 5.09 \end{bmatrix}, \quad u_s(t) \sim N\left(0, \sigma_2^2\right), \sigma_2 = 1.8$$
$$\varepsilon_s(t) = 0.27\varepsilon_s(t-1) + 0.46\varepsilon_s(t-2) - 0.04\varepsilon_s(t-3) + 0.10\varepsilon_s(t-4) + u_s(t) + 0.74u_s(t-1) \tag{18}$$

Table 7
Average MAE, RMSE, Type I and Type II error rates across measurement sites in group V for NLMM.

	C^{*}	C^{**}	C^{***}	C_a^{*}	C_a^{**}	C_a^{***}	M^{*}	M^{**}
$\overline{\text{MAE}}$ [°C]	2.74	2.12	2.55	2.55	1.91	2.22	2.24	2.73
$\overline{\text{RMSE}}$[°C]	3.20	2.54	2.98	2.95	2.28	2.61	2.65	3.17
$\tilde{E}_1$	0.15	0.45	0.26	0.20	0.34	0.22	0.09	0.11
$\tilde{E}_2$	0.11	0.04	0.12	0.10	0.06	0.10	0.12	0.16

Finally, Eqs. (19)–(20) present M^{*},and M^{**}which were based on the logistic air-to-water association profiles proposed in Mohseni et al. (1998):

$$M^{*}: \widehat{T_{w,s}}(t) = 5.61 + d_{1,s} + \frac{25.91 + d_{2,s} - 5.61 - d_{1,s}}{\exp\left\{2.72 + d_{3,s} - 0.19T_{a,s}(t)\right\}} + \varepsilon_s(t)$$
$$\Psi_{3,1} = \begin{bmatrix} 2.37 & 0.56 & 0.68 \\ 0.56 & 2.71 & 0.61 \\ 0.68 & 0.61 & 0.46 \end{bmatrix}, \quad \sigma = 1.96 \tag{19}$$

In M^{**}, average and maximum annual air temperatures in the calibration period explained site-specific variability related respectively to the 'floors' and the 'ceilings' of the air-to-water association profiles:

$$M^{**}: \widehat{T_{w,s}}(t) = 0.39X_{1,s} + d_{1,s} + \frac{0.68X_{2,s} + d_{2,s} - 0.39X_{1,s} - d_{1,s}}{\exp\left\{2.72 + d_{3,s} - 0.19T_{a,s}(t)\right\}} + \varepsilon_s(t)$$
$$\Psi_{3,2} = \begin{bmatrix} 2.00 & 0.38 & 0.53 \\ 0.38 & 2.11 & 0.58 \\ 0.53 & 0.58 & 0.45 \end{bmatrix}, \quad \sigma = 1.96 \tag{20}$$

Notably, residual standard deviation in (19) and (20) was substantially higher relative to models based on estimated annual cycles; hence one may expect inferior forecasting performance from these models. M^{***} is not presented as the NLMM estimation algorithm (Pinheiro and Bates, 1995) failed to converge for all examined ARMA specifications.

3.5. Mixed models: forecasting performance

Figs. 5–6 and Table 7 summarize the forecasting performance of NLMM. Before commenting on the results, we should stress the analogies between NLMM and their site-specific counterparts. C^{*}, $C^{**}$and C^{***}, which were based on C, did not use information on air temperatures, so are expected to display weaker forecasting performance relative to C_a^{*}, $C_a^{**}$and C_a^{***}, respectively. On the other hand, the latter specifications can be viewed as 1-day ahead forecasting models whereas the former can be used to forecast temperatures even at an annual horizon, since they only use information on day of the year, geographical and environmental site-specific variables. Table 7 suggests that using air temperatures in NLMM formulations based on seasonal decompositions reduced MAE and RMSE by about 0.2 °C; this reduction was substantially smaller relative to that depicted in Table 3 for site-specific models C and C_a.

Parameters in C^{*}, C_a^{*},and M^{*} were fixed across measurement sites whereas some parameters in C^{**}, C^{***}, C_a^{**}, C_a^{***} and M^{**} depended on site-specific maximum and mean annual air temperatures. This additional flexibility resulted into reductions in MAE and RMSE, which range from 0.2 °C to 0.6 °C for models based on annual components. The opposite was observed for models based on Mohseni's et al. (1998) approach: NLMMs that used information on air temperatures to predict 'floors' and 'ceilings' in air-to-water association profiles performed significantly worse than their 'static' counterparts according to all metrics.

As expected, performance of mixed models was inferior relative to site-specific models. For instance, average MAE increased by about 1 °C for $C^{**}$and C_a^{**} relative to C and C_a respectively (Tables 3 and 6). Fig. 6 depicts observed versus forecasted values for C_a^{**} and M^{**} for 4 sites in group V; the superiority of the former specification can be clearly observed. C_a^{**}outperformed alternative specifications according to average and median (across sites in group V) MAE and RMSE, and displayed significantly less bias and low error rates related to over-prediction of high-temperatures (Type II error).

Median and average MAE and RMSE for C_a^{**} were close to 2 °C, which was only about 0.5 °C higher relative to the performance of its site-specific counterparts reported in Table 3 and in Ahmadi-Nedushan et al. (2007). C_a^{**} was the best performing model in terms of MAE and RMSE in half of the locations in group V; furthermore, it was among the best three models in all but one locations. Accuracy of the best performing NLMM was found comparable to the one achieved from linear site-specific models (Table 3); this is a promising result, especially given that forecasting performance can be improved with estimation based on more measurement sites and more site-specific environmental and geographical variables. On the other hand NLMM based on air-to-water temperature association profiles, were found superior in terms of Type I error rates (Fig. 5d).

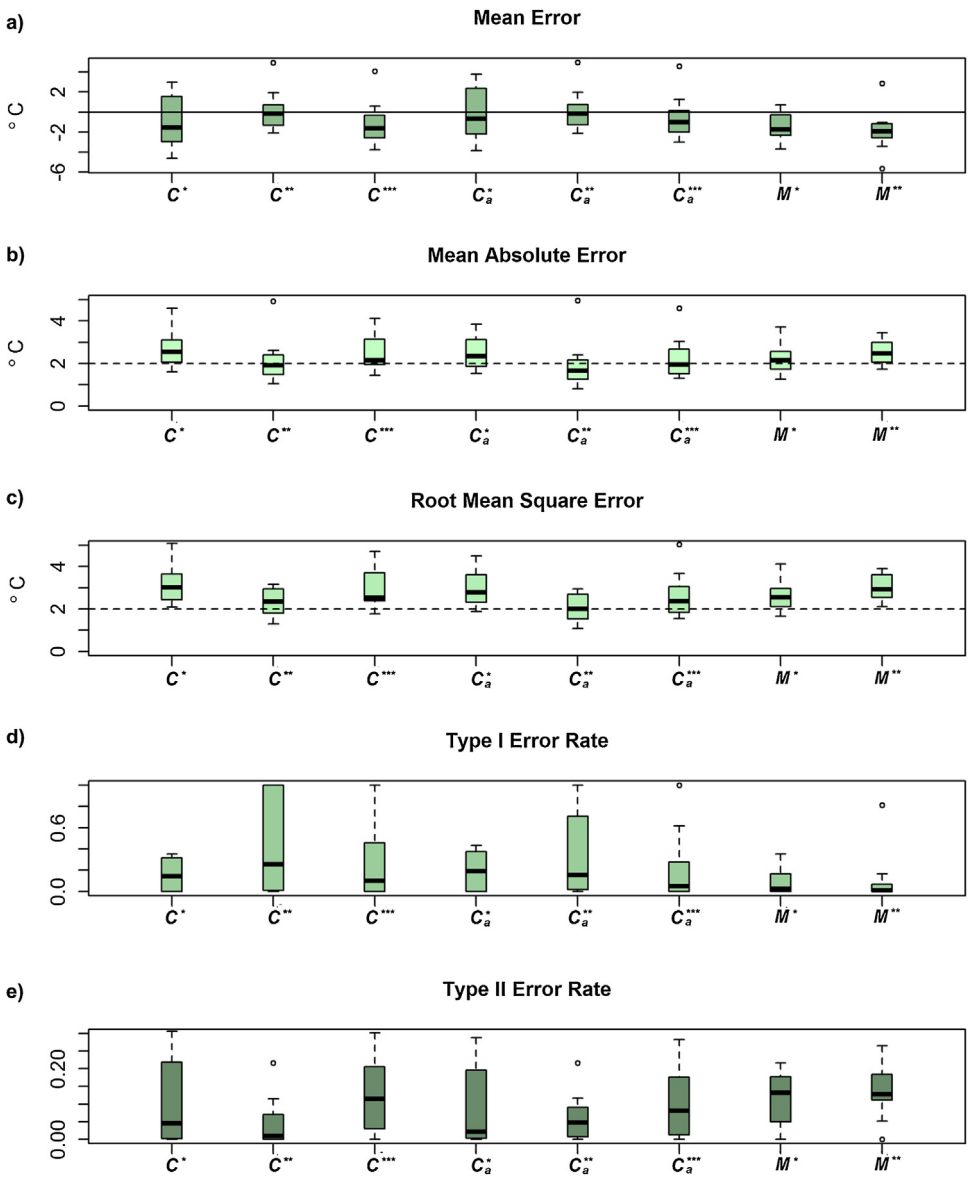

Fig. 5. Distributions of (a) mean error (bias); (b) MAE; (c) RMSE; (d) Type I and (e) Type II error rate across measurement sites in group V for alternative NLMM specifications.

4. Conclusions

This article applied a new model class, namely nonlinear mixed models, which makes forecasts of river water temperatures feasible in regions with insufficient data for site-specific model estimation. Forecasting accuracy was found comparable to the one achieved from linear site-specific time-series models. Therefore, the proposed approach is a promising tool that may extend the scope of statistical river temperature forecasting models.

Site-specific time-series models were used as the basis of NLMM building. The pool of specifications encompassed simple linear regressions (as in Arismendi et al., 2014), seasonal decompositions of temperatures coupled with time-series regressions for residual dynamics (as in Caissie et al., 2001; Hague and Patterson, 2014) and logistic air-to-water temperature association profiles, originally developed for weekly data by Mohseni et al. (1998). An evaluation of the dependence of site-specific model parameters on geographic and environmental factors revealed statistically significant relationships. These findings were exploited in the development of mixed models; a significant amount of variability in random effects, which correspond to site-specific parameters, was explained by site-specific characteristics.

The article focused on the Iberian Peninsula: a region with a high diversity of freshwater systems which lacked studies related to river water temperatures as opposed to Central (Switzerland, France) or Northern Europe (United Kingdom,

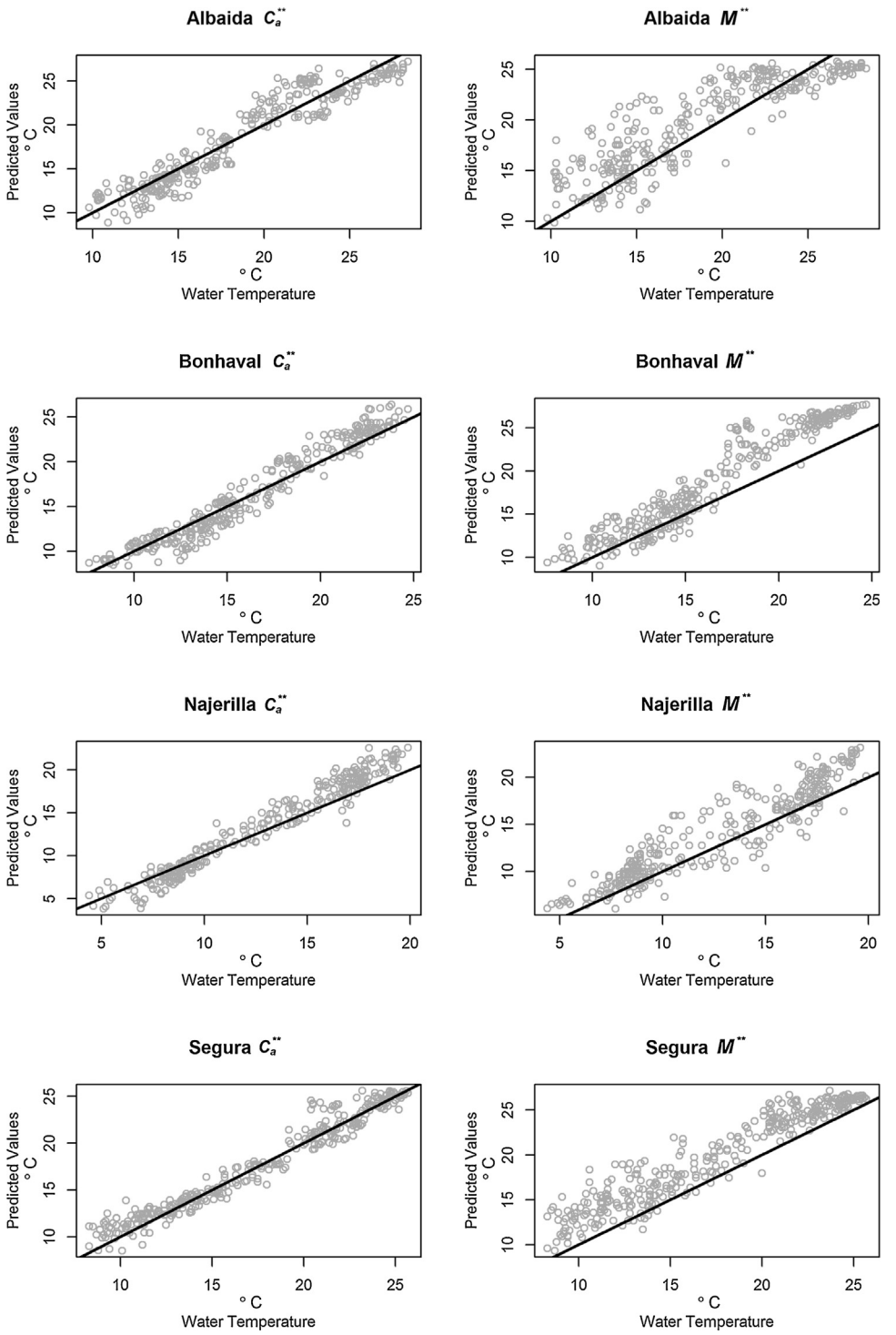

Fig. 6. Observed versus forecasted water temperatures from models C_a^{**} (left) and M^{**} (right) for four measurement locations in group V. The black line indicates where perfect forecasts should lie.

Sweden). A next step would be to apply similar analyses to other areas in the Mediterranean because the capacity of its inland water systems to act as sensors of global and local changes is particularly important. Such studies will improve the forecasting accuracy of river water temperatures, providing a tool for policy makers to facilitate strategic decisions and to manage freshwater resources according to water policy directives such as the EU-WFD.

Conflict of interest

None.

Acknowledgements

Yiannis Kamarianakis was partially supported by the National Science Foundation under Award DMS-1419593. The authors would like to thank the Spanish Environmental Department for providing the data, and three reviewers and the editor for their constructive comments. The analyzed datasets will be made freely available to download from the first author's website.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ejrh.2016.01.003.

References

Ahmadi-Nedushan, B., St-Hilaire, A., Ouarda, T.B.M.J., Bilodeau, L., Robichaud, E., Thiemonge, N., Bobee, B., 2007. Predicting river water temperatures using stochastic models: case study of the Moisie River. Hydrol. Processes 21, 21–34.
Allan, J.D., 1995. Stream Ecology: Structure and Function of Running Waters. Kluwer Dordrecht, The Netherlands.
Arismendi, I., Safeeq, M., Dunham, J.B., Johnson, S.L., 2014. Can air temperature be used to project influences of climate change on stream temperature? Environ. Res. Lett. 9, 084015.
Benyahya, L., Caissie, D., St-Hilaire, A., Ouarda, T.B.M.J., Bobee, B., 2007. A review of statistical water temperature models. Can. Water Resour. J. 32 (3), 179–192.
Benyahya, L., Caissie, D., El-Jabi, N., Satish, M.G., 2010. Comparison of microclimate vs. remote meteorological data and results applied to a water temperature model (Miramichi River, Canada). J. Hydrol. 380, 247–259.
Breusch, T.S., Pagan, A.R., 1979. A simple test for heteroscedasticity and random coefficient variation. Econometrica 47, 1287–1294.
Brey, T., 2010. An empirical model for estimating aquatic invertebrate respiration. Methods Ecol. Evol. 1 (1), 92–101.
Caissie, D., 2006. The thermal regime of rivers: a review. Freshwater Biol. 51 (8), 1389–1406.
Caissie, D., El-Jabi, N., Satish, M.G., 2001. Modelling of maximum daily water temperatures in a small stream using air temperatures. J. Hydrol. 251, 14–28.
Caissie, D., El-Jabi, N., St-Hilaire, A., 1998. Stochastic modeling of water temperatures in a small stream using air to water relations. Can. J. Civil Eng. 25, 250–260.
Caissie, D., Satish, M.G., El-Jabi, N., 2007. Predicting water temperatures using a deterministic model: application on Miramichi river catchments (New Brunswick, Canada). J. Hydrol. 336, 303–315.
Cluis, D., 1972. Relationship between stream water temperature and ambient air temperature a simple autoregressive model for mean daily stream water temperature fluctuations. Nordic Hydrol. 3 (2), 65–71.
Daigle, A., St-Hilaire, A., Peters, D., Baird, D., 2010. Multivariate modelling of water temperature in the Okanagan Watershed. Can. Water Resour. J. 35 (3), 237–258.
Development Core Team, R, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.
Dielman, T.E., 1986. A comparison of forecasts from least absolute value and least squares regression. J. Forecasting 5 (3), 189–195.
Godfrey, L.G., 1978. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. Econometrica 46, 1293–1302.
Guillemette, N., St-Hilaire, A., Ouarda, T.B.M.J., Bergeron, N., Robichaud, E., Bilodeau, L., 2009. Feasibility study of a geostatistical modelling of monthly maximum stream temperatures in a multivariate space. J. Hydrol. 364, 1–12.
Gustafson, M., 2008. Effects of thermal regime on mayfly assemblages in mountain streams. Hydrobiologia 605, 235–246.
Hague, M.J., Patterson, D.A., 2014. Evaluation of statistical river temperature forecast models for fisheries management. North Am. J. Fish. Manage. 34, 132–146.
IPCC, Working Group I Contribution to the IPCC Fifth Assessment Report, 2013. Climate Change 2013: The Physical Sciences Basis Summary for Policymakers. Cambridge University Press.
Isaak, D.J., Wollrab, S., Horan, D., Chandler, G., 2010. Climate change effects on stream and river temperatures across the northwest U.S. from 1980 to 2009 and implications for salmonid fishes. Clim. Change 113 (2), 499–524.
Jeong, D.I., Daigle, A., St-Hilaire, A., 2013. Development of a stochastic water temperature model and projection of future water temperature and extreme events in the Quelle river basin in Quebec, Canada. River Res. Appl. 29, 805–821.
Hyndman, R.J., Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for R. J. Stat. Software 27 (3), 1–22.
Keenan, D.M., 1985. A Tukey nonadditivity-type test for time series nonlinearity. Biometrika 72, 39–44.
Koenker, R., Park, B.J., 1994. An interior point algorithm for nonlinear quantile regression. J. Econometrics 71 (1–2), 265–283.
Larnier, K., Roux, H., Dartus, D., Croze, O., 2010. Water temperature modeling in the Garonne River (France). Knowl. Manage. Aquat. Ecosyst. 398, 1–20.
Lessard, J.L., Hayes, D.B., 2003. Effects of elevated water temperature on fish and macroinvertebrate communities below small dams. River Res. Appl. 19 (7), 721–732.
Ljung, G.M., Box, G.E.P., 1978. On a measure of lack-of-fit in time series models. Biometrika 65 (2), 297–303.
Mantua, N., Tohver, I., Hamlet, A., 2010. Climate change impacts on streamflow extremes and summertime stream temperature and their possible consequences for freshwater salmon habitat in Washington State. Clim. Change 102 (1–2), 187–223.
MIMAM (Ministerio de Medio Ambiente, Spanish Ministry of Environment), 2004. Water in Spain (Spanish Water White Paper), Madrid,I.S.B.N. 84-8320-219-0.
Mohseni, O., Stefan, H.G., Eaton, J.G., 2003. Global warming and potential changes in fish habitat in US streams. Clim. Change 59 (3), 389–409.
Mohseni, O., Stefan, H.G., Erickson, T.R., 1998. A nonlinear regression model for weekly stream temperatures. Water Resour. Res. 34 (10), 2685–2692.
Pappenberger, F., Ramos, M.H., Cloke, H.L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., Salamon, P., 2015. How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. J. Hydrol. 522, 697–713.
Pinheiro, J.C., Bates, D.M., 1995. Approximations to the log-likelihood function in the nonlinear mixed-effects model. J. Comput. Graphical Stat. 4 (1), 12–35.
Pinheiro, J.C., Bates, D.M., 2009. Mixed Effects Models in S and S-plus. Springer.
Rall, B.C., Vucic-Pestic, O., Ehnes, R.B., Emmerson, M., Brose, U., 2010. Temperature, predator–prey interaction strength and population stability. Global Change Biol. 16 (8), 2145–2157.

Stefan, H.G., Sinokrot, B.A., 1993. Projected global climate change impact on water temperatures in five north central US streams. Clim. Change 24 (4), 353–381.
Thuiller, W., Lavergne, S., Roquet, C., Boulangeat, I., Lafourcade, B., Araujo, M.B., 2011. Consequences of climate change on the tree of life in Europe. Nature 470, 531–534.
Tsay, R.S., 1986. Nonlinearity test for time series. Biometrika 73, 461–466.
Wenger, S.J., Isaak, D.J., Luce, C.H., Neville, H.M., Fausch, K.D., Dunham, J.B., Dauwalter, D.C., Young, M.K., Elsner, M.M., Rieman, B.E., Hamlet, A.F., Williams, J.E., 2011. Flow regime, temperature, and biotic interactions drive differential declines of trout species under climate change. PNAS 108 (34), 14175–14180.
Zuur, A., Ieno, E.N., Walker, N., Saveliev, A.A., Smith, G.M., 2009. Mixed Effects Models and Extensions in Ecology with R. Springer.