Evaluating the Performance of Hydrological Models via Cross-Spectral Analysis: Case Study of the Thames Basin, United Kingdom

GRAHAM P. WEEDON

Met Office, Joint Centre for Hydro-Meteorological Research, Wallingford, United Kingdom

CHRISTEL PRUDHOMME, SUE CROOKS, RICHARD J. ELLIS, AND SONJA S. FOLWELL

Centre for Ecology and Hydrology, Wallingford, United Kingdom

MARTIN J. BEST

Met Office, Exeter, United Kingdom

(Manuscript received 20 January 2014, in final form 17 October 2014)

ABSTRACT

Nine distributed hydrological models, forced with common meteorological inputs, simulated naturalized daily discharge from the Thames basin for 1963–2001. While model-dependent evaporative losses are critical for modeling mean discharge, multiple physical processes at many time scales influence the variability and timing of discharge. Here the use of cross-spectral analysis is advocated to measure how the average amplitude—and independently, the average phase—of modeled discharge differ from observed discharge at daily to decadal time scales. Simulation of the spectral properties of the model discharge via numerical manipulation of precipitation confirms that modeled transformation involves runoff generation and routing that amplify the annual cycle, while subsurface storage and routing of runoff between grid boxes introduces most of the autocorrelation and delays. Too much or too little modeled evaporation affects discharge variability, as do the capacity and time constants of modeled stores. Additionally, the performance of specific models would improve if four issues were tackled: 1) nonsinusoidal annual variations in model discharge (prolonged low base flow and shortened high base flow; three models), 2) excessive attenuation of high-frequency variability (three models), 3) excessive short-term variability in winter half years but too little variability in summer half years (two models), and 4) introduction of phase delays at the annual scale only during runoff generation (three models) or only during routing (one model). Cross-spectral analysis reveals how reruns of one model using alternative methods of runoff generation—designed to improve performance at the weekly to monthly time scales—degraded performance at the annual scale. The cross-spectral approach facilitates hydrological model diagnoses and development.

1. Introduction

Within the Water and Global Change (WATCH) European Union (EU) Sixth Framework Programme (FP6), a variety of distributed hydrological models were run globally, excluding the effects of anthropogenic land cover and management. Additionally, river basin models were run for specific basins. The Water Model Intercomparison Project (WaterMIP) protocol adopted during WATCH (Haddeland et al. 2011) used common meteorological forcing data for the twentieth century provided at 0.5° × 0.5° resolution (Weedon et al. 2011), a common routing network for surface and subsurface runoff between grid boxes, and a common reporting format (www.eu-watch.org/watermipprotocol2009a). This paper presents a practical approach to quantitative comparison of daily discharge outputs from WATCH models, and hence allows analysis of model performance, as demonstrated for the Thames basin of southeast England, United Kingdom.

Typically, metrics of hydrological model performance for comparing observations with model output include correlation, root-mean-square error (RMSE), mean bias

Corresponding author address: Graham Weedon, Met Office, Joint Centre for Hydro-Meteorological Research, Maclean Building, Crowmarsh Gifford, Wallingford, Oxfordshire OX10 8BB, United Kingdom.
E-mail: graham.weedon@metoffice.gov.uk

DOI: 10.1175/JHM-D-14-0021.1

error (MBE), and standard deviation or related indices such as Nash–Sutcliffe efficiency (NSE; Nash and Sutcliffe 1970). When assessing several climate models, Taylor diagrams, based on standard deviation and correlation, are often used to combine multiple metrics visually, allowing one to assess whether a particular model outperforms others and/or whether model developments are leading to improved performance (Taylor 2001; Gleckler et al. 2008).

In the context of modeling very large basins, domination of the discharge variability by a small range of frequencies (e.g., cycles at the annual scale) means that analysis with traditional metrics is not a problem and cross-spectral methods are not needed. However, in general, with the exception of MBE, these metrics often fail to provide sufficiently unambiguous insights into the ways in which particular model outputs differ from observations or from other models (Lane 2007). This is because, for example, day-to-day variability and persistence of high and low discharge are not measured separately because of the averaging across all time scales. However, by measuring the relative difference of modeled from observed discharge, MBE can be related directly to the two main factors influencing the water balance in a catchment: the inputs (in the form of precipitation) and the residual from the losses (through evapotranspiration) over multiple years or decades (assuming subsurface storage is approximately constant in the long term). Here, MBE is used with modeled discharge versus observed naturalized discharge.

Very often, visual inspection of modeled and observed discharge time series is extremely informative as this reveals the variability at all time scales. For example, this might show that the RMSE and standard deviation would be improved by increasing the magnitude of the short-term responsiveness of a hydrological model to precipitation events or that correlation would be increased by improving the timing of the annual/seasonal cycle of the modeled discharge relative to observations. Expert hydrologists often use visual inspection of the hydrographs in their assessment of a modeled response, but this can be subjective and is not a quantified (objective) measure.

Here, we investigate how a cross-spectral approach to comparing model outputs with observations can yield physical insights into the behavior and deficiencies in models. The approach advocated here differs from, but is complementary to, a more traditional approach in hydrology where spectral properties are used numerically to help refine model parameter estimates (e.g., Montanari and Toth 2007; Quets et al. 2010; Pauwels and De Lannoy 2011; Moussu et al. 2011). However, by combining the mismatches across all frequencies into their objective functions, the traditional approach does not consider frequency-specific model deficiencies.

We are not primarily concerned with identifying optimal model parameters and structures, but rather with quantifying model deficiencies, by analyzing amplitudes and phases at different time scales, to concentrate on how specific physical processes are represented. This analysis concerns average model performance rather than alternative methods that concentrate specifically on extreme high or low flows. We focus on average mismatches for 1963–2001 rather than localizing in time when the specific mismatches between model outputs and observations have occurred [compare use of wavelets (e.g., Smith et al. 1998; Labat et al. 2000b; Lane 2007; Schaefli and Zehe 2009; Labat 2010; Liu et al. 2011) and other approaches (e.g., Herman et al. 2013)].

Padilla and Pulido-Bosch (1995) used cross-spectral analysis for comparing discharge with precipitation in Spanish and French karst systems. However, Labat et al. (2000a,b) showed that karst systems can be so physically heterogeneous and dynamically varying that the discharge variability is not sufficiently characterized by the averaging inherent in the Fourier methods discussed here. By contrast, the Thames basin (and many other basins; Milly and Wetherald 2002) has a far simpler geometry, far longer response times (Naden 1992), and little input from snowmelt events, so the power spectra and cross spectra of the precipitation and discharge data provide meaningful estimates of the average variability at different time scales.

After introducing the Thames basin and observations (section 2) and models (section 3), we outline the methods (section 4) with technical details in the appendixes. We consider the transformation of the observed precipitation to observed discharge (section 5a) as well as the transformation of the gridded precipitation to modeled discharge (section 5b). The runoff generation in separate grid boxes and routing mechanisms in the distributed models are discussed in section 5c. We illustrate how we can reproduce the key spectral characteristics of the modeled runoff and discharge outputs by simple numerical manipulation of gridbox precipitation (section 5d). Understanding the origin of the spectral properties then allows evaluation of the models via amplitude ratio and phase spectra for comparing the modeled with the observed discharge time series (section 6).

2. Data from the Thames basin

In southeast England, the Thames basin to Kingston, the lowest gauging station on the river and a short distance upstream of its tidal limit, covers an area of about 9947 km^2 with relatively subdued topography (maximum

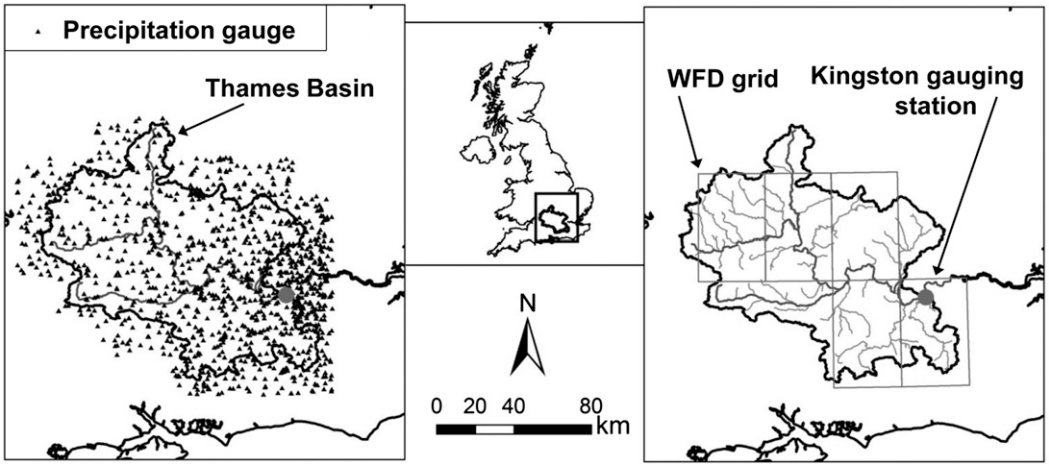

FIG. 1. (middle) Location of the Thames basin within the United Kingdom. Scale bar relates to maps in (left) and (right). (left) The 1265 daily precipitation gauges used to generate the observed basinwide-average precipitation for 1963–2001. The gauges used extend out to 1.5 km beyond the limit of the CLASSIC 20-km grid (not shown). (right) Five regular 0.5° lat–lon grid boxes from the WFD grid are connected within the DDM30 routing network used in WATCH and WaterMIP. Basinwide WFD precipitation was obtained as an unweighted average of the daily WFD precipitation in the five grid boxes shown. Model discharge was obtained from the grid box containing the Kingston discharge gauging station (indicated by the gray circle).

elevation less than 330 m) and an average (for 1961–90) of 720 mm of precipitation per year (Fig. 1). On average there are only a few millimeters of snowmelt per year, but substantial variations from year to year; generally, snowmelt has negligible impact on the timing of river flow. Around 65% of the precipitation, which occurs year-round, is lost to evaporation, especially from April to September.

A daily time series of the area-average precipitation in the Thames basin was derived using the triangle method of Jones (1983) from, on any particular day, an average of around 1000 out of 1265 daily rain gauges available for 1963–2001 (Fig. 1). Since 1986, discharge is measured at Kingston using a multipath ultrasonic gauge, but from 1974 to 1986 it was assessed via a single-path ultrasonic gauge. Earlier than this, the flow was measured at the tidal limit (at Teddington, about 2 km downstream from Kingston) from a complex system of gates, sluices, and weirs, with a tail water rating between level and discharge. A naturalized daily discharge record is available where the gauged flows have been adjusted for the net impact of upstream abstractions and discharges (Marsh and Hannaford 2008); this has been used as the observed naturalized discharge series.

The Thames basin is diverse in terms of geology, with 45% covered by permeable rocks and providing substantial groundwater flow. The remainder of the basin is characterized by more responsive flow from less permeable soils, particularly clays. The subannual-frequency ranges of the spectra described later are subdivided using response times characteristic of different physical processes. For the permeable parts of the catchment, a response time period of 2–6 months (60–182.5 days) is typical, that is, the slow response (SR) scale. For more responsive areas, the longest response (analogous to concentration time) is around 7 days. As two flood peaks are considered independent if separated by 3 times the catchment response time, an interval of 7–21 days defines the quick response (QR) scale. Water takes 2–4 days to flow directly from impermeable surfaces and down the channel from the headwaters to the discharge measurement point—the surface runoff and channel routing (SCR) scale.

3. Meteorological forcing data and models

The WATCH models were forced using the WATCH forcing data (WFD) that are based on the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40), which was interpolated, elevation corrected, and adjusted at the monthly scale to match gridded observations (Weedon et al. 2011). The WFD provide 0.5°-spatial-resolution, three-hourly data for near-surface air temperature, wind speed, pressure, specific humidity, downward longwave radiation flux, downward shortwave radiation flux, rainfall rate, and snowfall rate.

Five WFD grid boxes cover the Thames basin as connected by the 0.5° 30-min global drainage direction map (DDM30) routing network for surface runoff plus subsurface runoff (Fig. 1; Döll and Lehner 2002). Modeled discharge was assessed using the grid box containing the discharge gauging station at Kingston. Grid boxes

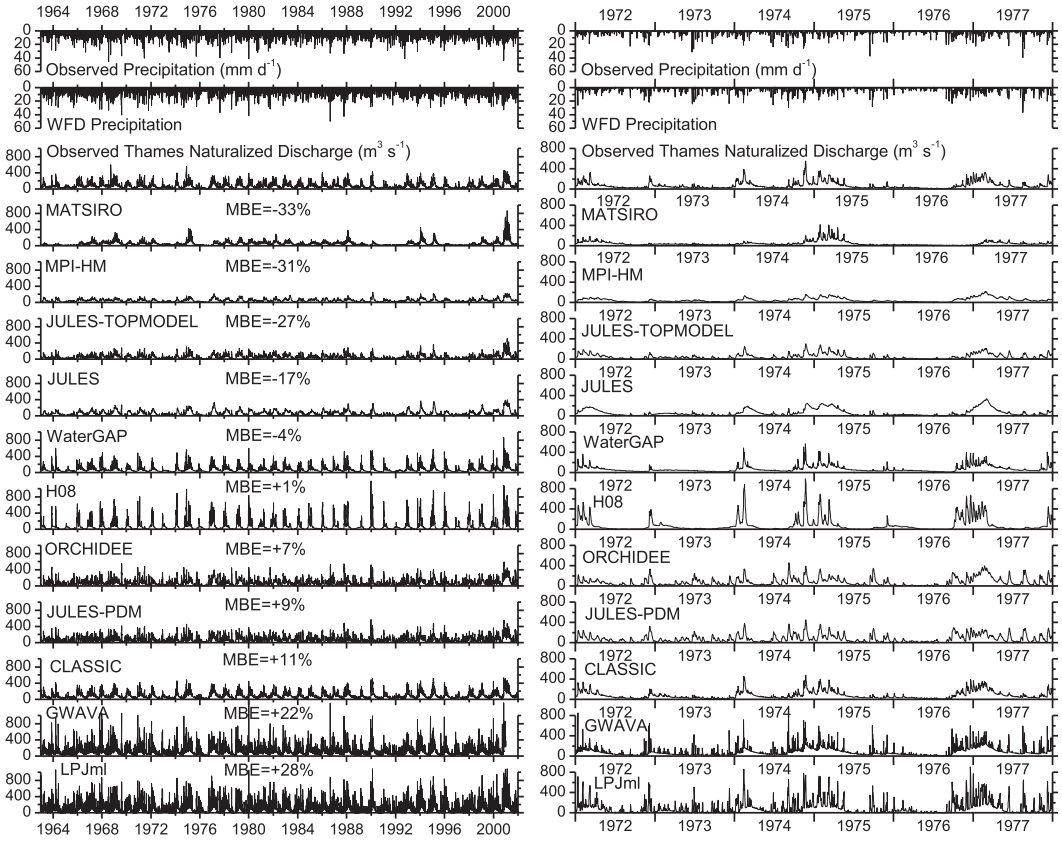

FIG. 2. (left) Time series for 1963–2001 of observed and WFD daily precipitation (top two) and of observed and modeled daily discharge (lower twelve) for the Thames basin. (right) Identical data, but for 1972–77 to allow visualization of subannual variability.

external to those shown in Fig. 1 are specified by the DDM30 network as draining outside the basin.

Comparing the daily WFD precipitation time series—averaged across the five grid boxes—with the observed precipitation (Fig. 2; $N = 14\,245$ days) shows similar means (mean ± 95% confidence interval for observed = 1.97 ± 0.06 mm versus WFD = 1.96 ± 0.06 mm), standard deviations (3.68 mm versus 3.78 mm), lag-1 autocorrelation (0.24 versus 0.23), and tolerable mean absolute error (MAE; 1.58 mm). These very similar results occur despite the land covered by the five grid boxes being about 1.5% less than the area defined by the catchment boundary. The highly significant Pearson's r correlation (0.63, $P < 0.001$) and correlation adjusted for lag-1 autocorrelation (0.54, $P < 0.001$; Ebisuzaki 1997) are not higher because of the limitations of ERA-40 in representing clouds (and hence precipitation events), especially associated with local convection (Weedon et al. 2011). The Global Precipitation Climatology Centre, version 4 (GPCCv4), gridded precipitation gauge totals incorporate some of the Thames basin gauge observations. These GPCCv4 totals were used in monthly bias correction of the WFD precipitation (Weedon et al. 2011), so there is an inherent similarity at monthly and longer time scales between the observed and WFD precipitation. The consistency of the observed and model-input precipitation series suggests that large biases in modeled discharge are not attributable to the use of WFD precipitation data in forcing the models.

The different hydrological models (see Table 1 for expansions), simulating unmanaged conditions for 1963–2001 (i.e., excluding water abstraction, irrigation, regulation, and human-related land cover changes), were either operated at 1) daily time steps for the river basin model (CLASSIC) and for the global hydrological models (MPI-HM, WaterGAP, GWAVA, and LPJmL) or 2) hourly or half-hourly time steps for the land surface models (MATSIRO, JULES, H08, and ORCHIDEE) to allow calculation of the diurnal energy balance.

We include all models from WATCH and WaterMIP that provided daily discharge time series for 1963–2001 (www.eu-watch.org/data_availability). Haddeland et al. (2011) provide summaries and references to the designs of the models (excluding CLASSIC) with comparisons of

TABLE 1. Metrics comparing modeled and observed daily naturalized discharge for 1963–2001 in the Thames basin. Variable N = 14 245, except for observed discharge vs GWAVA discharge using 1963–2000, where N = 13 880. The mean absolute error provides the average absolute difference (Willmott and Matsuura 2005) of individual modeled daily discharge values from observed values. Pearson's r recalculated after prewhitening derived using the values of lag-1 autocorrelation (Ebisuzaki 1997) is given by r_{Adj}. All the correlation and adjusted correlation values listed are highly significant ($P < 0.001$)—note the high N.

Observations or model	Mean ($m^3 s^{-1}$)	MBE (% ± 95% CI)	Std dev ($m^3 s^{-1}$)	MAE ($m^3 s^{-1}$)	NSE	r	ρ_1	r_{Adj}
Observations	82.54	—	66.63	—	—	—	0.960	—
Minimal Advanced Treatments of Surface Interaction and Runoff (MATSIRO)	55.34	−32.96 ± 6.58	50.43	34.20	+0.251	+0.654	0.957	+0.387
Max Planck Institute–Hydrology Model (MPI-HM)	57.12	−30.79 ± 2.53	36.90	29.43	+0.431	+0.797	0.996	+0.453
JULES–Topography-Based Model (JULES-TOPMODEL)	60.21	−27.05 ± 3.74	55.49	28.26	+0.661	+0.881	0.986	+0.542
Joint UK Land Environment Simulator (JULES)	68.84	−16.60 ± 4.02	63.94	30.62	+0.530	+0.778	0.994	+0.430
Water–Global Assessment and Prognosis (WaterGAP)	79.08	−4.20 ± 3.44	62.82	19.89	+0.722	+0.856	0.904	+0.278
Hanasaki et al. (2008) model (H08)	83.08	+0.65 ± 7.88	127.32	48.83	−0.544	+0.813	0.977	+0.474
Organizing Carbon and Hydrology in Dynamic Ecosystems (ORCHIDEE)	88.69	+7.44 ± 4.98	76.48	39.05	+0.210	+0.669	0.974	+0.422
JULES–Probability Distributed Moisture (JULES-PDM)	89.52	+8.45 ± 4.84	76.26	37.41	+0.382	+0.744	0.977	+0.528
Climate and Land-Use Scenario Simulation in Catchments (CLASSIC)	91.58	+10.95 ± 3.92	72.83	18.84	+0.804	+0.923	0.957	+0.565
Global Water Availability Assessment model (GWAVA)	99.24	+21.86 ± 4.70	94.29	37.36	−0.417	+0.601	0.447	−0.181
Lund–Potsdam–Jena managed Land (LPJmL)	105.55	+27.87 ± 4.55	128.46	56.91	−1.049	+0.723	0.827	+0.483

monthly average outputs globally and for selected large catchments. CLASSIC uses a 20 km × 20 km grid, rather than the WATCH 0.5° grid, with flow paths and runoff delays represented as a kinematic wave from headwater grid boxes to the outlet grid box (Crooks and Naden 2007).

The only models calibrated using local measured streamflow data are CLASSIC and WaterGAP. Within WATCH, GWAVA was redesigned to run on daily rather than monthly time steps, but daily routing was not implemented, so daily runoff was rescaled and aggregated to produce the daily discharge values (corrected and reuploaded since the WaterMIP study, but not available for 2001). WaterGAP and LPJmL did not provide discharge values for leap days, so the averages of values from 28 February and 1 March were used, affecting just 10 out of 14 245 days (1963–2001). The differences between models illustrated here for the Thames basin will vary elsewhere since model performance is also linked to catchment properties and regional meteorology (Gudmundsson et al. 2012a,b).

In the standard JULES run for WATCH, runoff generation occurs with Darcian drainage without subgridbox heterogeneity (Best et al. 2011). JULES was rerun, using the overall configuration employed in WATCH, to explore whether implementation of alternative conceptual runoff generation process methods would improve the partition between surface and subsurface/groundwater (important for the Thames basin given the wide range of bedrock permeability). JULES-TOPMODEL (e.g., Beven et al. 1984) uses a within-grid probability distribution of soil saturation and water tables (Gedney and Cox 2003; Clark and Gedney 2008). JULES-PDM uses the PDM soil moisture method to account for within-grid soil heterogeneity and saturation excess runoff via a probability distribution of soil stores (Moore and Clarke 1981; Clark and Gedney 2008).

Figure 2 illustrates the observed and modeled discharge below the average basinwide precipitation time series. Figure 2 (right) allows visualization of subannual variability during 1972–77, including the major hydrological drought of 1975/76. Table 1 provides a comparison of the modeled with the observed naturalized daily discharge using standard metrics ordered by increasing MBE.

MBE indicates the average modeled discharge minus the average observed discharge divided by average observed discharge, as a percentage. The 95% confidence interval was derived using a Student's t value after calculating MBE separately for each calendar year (i.e., 39 values). Negative MBE in Table 1 denotes too little discharge on average, especially due to too much evaporation (vice versa for positive MBE). MBE differences relate to the way evaporation is calculated and to the

factors determining evaporative losses. In this case, the difference in area of the grid boxes draining to Kingston and the actual basin is too small to contribute substantially to the model MBEs. All models remove water from the soil through bare soil evaporation and/or evapotranspiration, but the soil is represented differently in terms of the capacity of stores and the control of the release of water to underlying levels and into the channel (i.e., in terms of both the partition between stores and the delays) and some models include direct evaporation of water intercepted by the canopy.

Mean absolute error rather than RMSE is provided in Table 1 given the problems with interpretation of the latter (Willmott and Matsuura 2005). Selecting metrics other than MBE would result in a different ordering of the models. Such potential reordering of models relates to the different dependency of the alternative metrics on mismatches in the amplitude of variation and/or mismatches in phase and the different time scales at which such mismatches occur.

4. Methods: Spectral and cross-spectral analysis

A time series is simply a time-ordered sequence of variable values (e.g., daily discharge). They are most importantly characterized in terms of the wavelength or period (i.e., the inverse of frequency) of the oscillations, the amplitude or deviation of the oscillations from the mean level, and the phase or timing of maxima and minima. According to Fourier's theorem, any time series containing oscillations, but no infinite values, can be decomposed into component sine and cosine waves via, for real data, for example, the discrete Fourier transform (DFT) to obtain the average amplitudes. The DFT is obtained by manipulating the data themselves so that with N data points and a sample rate of Δt, the frequency range of the spectrum is evaluated at $N/2 + 1$ locations between lowest ($=1/N\Delta t$) and highest, or Nyquist, frequency ($=1/2\Delta t$; appendix A). The periodogram shows the sum of the squared sine amplitude plus the squared cosine amplitude at each frequency, but a smoothed version, the estimated power spectrum, provides a better approximation to the expected result [for background, see, e.g., Priestley (1981), von Storch and Zwiers (1999), and Weedon (2003)].

The estimated power spectrum of a finite time series of pure random numbers (with zero serial or autocorrelation, white noise) has a horizontal background. An estimated power spectrum sloping down to the Nyquist frequency derives from red noise. A red noise spectrum that is linear on a Log(power)–Log(frequency) plot conforms to a power law, but if curving toward horizontal at the lowest and highest frequencies, it is typically associated with a lag-1 autocorrelation ρ_1 between zero and one indicative of first-order autoregression AR(1). Almost perfectly regular or quasi periodic processes cause concentrations of variance in narrow bands, creating power spectral peaks emerging from the background.

For comparing time series of the terrestrial flux of water out of the basin (output or discharge) with the flux of water from the atmosphere (input or precipitation), we use cross-spectral analysis (Padilla and Pulido-Bosch 1995): specifically, the gain spectrum and, closely related, the amplitude ratio spectrum, plus the phase spectrum (appendix A). This corresponds to analysis of the frequency response function or the spectral transfer function (Priestley 1981). Graphs of Log(gain)–Log(frequency) and (linear) phase–Log (frequency) are known as Bode plots and are used widely in systems analysis and control (Jenkins and Watts 1969). A gain or amplitude ratio exceeding one indicates amplification and less than one indicates attenuation. Here, a positive phase indicates that discharge variations lag precipitation variations (negative phase is physically impossible or noncausal).

Strictly, for Bode plots to be a complete description of the average system behavior, the system should be linear and time invariant without feedbacks (Jenkins and Watts 1969; Priestley 1981; Ifeachor and Jervis 1993). Characteristically, nonlinear systems generate harmonic spectral peaks at integer multiples of the frequency of primary input signals and sometimes combination tone peaks due to intermodulation between pairs of primary signals. Such frequency interactions require analysis with generalized frequency response functions rather than Bode plots (Billings 2013). However, the observed discharge spectrum of the Thames basin has no harmonic peaks associated with the annual cycle peak (section 5a); hence, it can be usefully analyzed with Bode plots (as implemented for many other basins; Padilla and Pulido-Bosch 1995; Milly and Wetherald 2002). Time invariance has been assumed; the precipitation and discharge time series are stationary in mean and variance and all analyses are for the same interval (1963–2001). Since the Thames causes, at most, minor inundation, there is essentially no feedback between discharge and precipitation.

Bode plots are used here to help interpret the processes involved in transforming precipitation into discharge variability. However, we evaluate average model performance by comparing two time series of the same variable—modeled discharge with observed discharge—by using amplitude ratio spectra and phase spectra (section 6). Unlike Bode plots, this requires no assumptions about the system being modeled. In this case, negative phase values are plausible, indicating that model discharge variations lead observed discharge variations.

5. Spectral characteristics of Thames basin precipitation, runoff, and discharge

a. *Observed discharge versus observed precipitation*

To allow spectral comparison of the observed discharge with the observed precipitation data, the latter were rescaled to discharge units (e.g., the average precipitation of 1.97 mm day^{-1} corresponds to 226.8 m^3 s^{-1}, while the average observed discharge is 82.5 m^3 s^{-1}). The power spectrum of the observed precipitation slopes gently to the right, with a modest spectral peak at the annual scale (significant at the 99.0% confidence level; Fig. 3b). The modest size of the annual spectral peak results from the relatively small seasonal variations in total observed precipitation (Fig. 2). The near-linear and low average slope of the observed precipitation spectrum in Fig. 3b is consistent with a small autocorrelation (ρ_1 about 0.2) and hence short-term memory characteristics (Kantelhardt et al. 2006).

The power spectrum of observed discharge (Fig. 3a) has a pronounced spectral peak reflecting large annual cycles in discharge (significant at the 99.999% level). The gain spectrum and amplitude-ratio spectrum in Fig. 3c show that in generating discharge the Thames basin attenuates precipitation variations at most frequencies, especially subannually. This contrasts with basins that are arid or have permafrost and/or substantial snowmelt when amplification is observed at multiple spectral background frequencies (Milly and Wetherald 2002). On the other hand, in the Thames basin, the gain and amplitude ratio exceed unity at the annual scale (amplitude ratio ±95% confidence interval = 1.37 +0.33/−0.27; Fig. 3c). In some karst basins, amplification is restricted to low frequencies (Padilla and Pulido-Bosch 1995), but amplification of precipitation variations at the annual scale is common in large humid midlatitude and tropical basins (Materia et al. 2010).

The phase spectrum shows that at the annual scale, the observed discharge variations are delayed by +75.6° ± 19.0° compared to the observed precipitation variations, or 75.6°/360° × 365.24 days = 76.7 days (Fig. 3d). The significance of the subannual-scale trend in the phase toward +180° at the Nyquist frequency is addressed later [section 5d(2)].

The shape of the background power spectrum of observed discharge is consistent with an AR(1) character (ρ_1 = 0.96; Table 1) and short-term memory. However, many studies have inferred a power-law character from the power spectra of monthly and annual discharge data from large basins (e.g., Pelletier and Turcotte 1997). Such an interpretation implies long-term memory associated with the Hurst phenomenon (Hurst 1951; Mesa and Poveda 1993; Heneghan and McDarby 2000; Schepers et al. 1992; Bryce and Sprague 2012; Fleming 2014).

Although there has been a lack of a physical explanation for the Hurst phenomenon (e.g., Mesa and Poveda 1993), Hoskins (1984) noted that aggregation of multiple independent short-term memory processes produces a power law. Mudelsee (2007) demonstrated via observations and modeling that the Hurst phenomenon arises progressively downstream because of the aggregation of discharge variations from separate tributaries. Fleming (2014), using annual observations of Thames discharge for 1883–2011, showed there were insufficient data to either demonstrate or rule out a power law. We invoke the explanation of the Hurst phenomenon by Mudelsee (2007) and interpret the lack of a power law in Fig. 3a—that extends down to daily frequencies—as resulting from the modest size of the Thames basin. The catchment area combines with the small elevation range and the consistent catchment response for different precipitation events (due to the dominance of the slow hillslope response; Naden 1992) so that the tributaries produce correlated, rather than independent, variations in water inputs.

Thus, we interpret the increased slope of the observed discharge spectrum compared to the observed precipitation spectrum as reflecting increased short-term memory (higher autocorrelation) caused by storage and routing processes (Milly and Wetherald 2002) rather than long-term memory processes. Critically, we believe this is more appropriate than using multiple power-law approximations for the spectral background (cf. Labat et al. 2000a).

b. *Modeled discharge versus WFD precipitation*

The cross-spectral relationship between observed discharge and observed precipitation is well reproduced by CLASSIC (Figs. 3e–h). Models other than CLASSIC show much less success with reproducing the amplitude variations of observed discharge; Figs. 3i–p show examples from near the extremes of the MBE-ordered models. Lack of variability in MPI-HM discharge (Fig. 2) leads to low gain and low-amplitude ratios compared to observations at all frequencies (Fig. 3k). Conversely, LPJmL has too much variability at most frequencies plus a drop or roll off in power near the Nyquist frequency not seen for the observations. The better performance of CLASSIC compared to the other models is likely to be the result of several factors, including that it is a river basin model, selected as appropriate for the basin being modeled, rather than a generic global model; it is calibrated; it is run at a higher resolution than the other models; and, uniquely to this study, it uses a kinematic wave approach to routing.

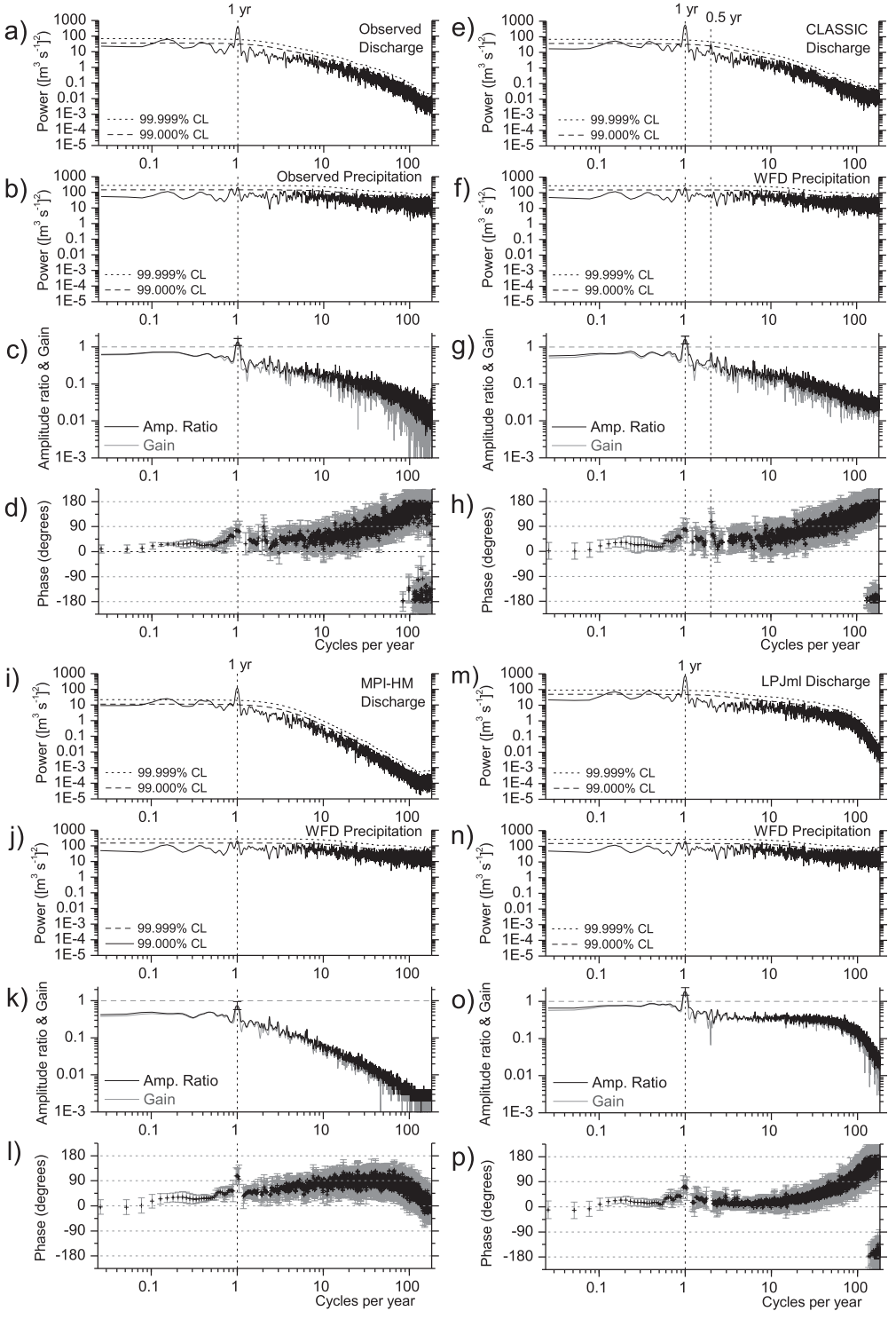

FIG. 3. (a) Power spectrum of observed daily discharge in the Thames basin. Dashed lines on the power spectra show the 99.0% and 99.999% confidence levels for detecting significant power spectral peaks. (b) Power spectrum of observed precipitation rescaled to $m^3 s^{-1}$. (c) Gain spectrum (gray) and amplitude ratio spectrum (black). The 95% confidence interval of amplitude ratio at the annual scale is shown using vertical black bar. (d) Phase spectrum, where the 95% confidence intervals are indicated by vertical gray bars. Phase values are only illustrated at frequencies where coherency exceeds to 95% level (i.e., 0.78), limiting phase 95% confidence intervals to $<\pm 51°$. (e)–(h) As in (a)–(d), but for CLASSIC-modeled discharge and the WFD precipitation basinwide data. (i),(j) Power spectra for MPI-HM–modeled discharge and WFD precipitation with (k),(l) corresponding gain, amplitude ratio, and phase spectra. (m)–(p) As in (i)–(l), but for discharge from LPJml.

The spectral peak at the half-annual scale in the spectrum of CLASSIC discharge represents a harmonic of the annual cycles, indicating slight nonlinearity in the model. CLASSIC simulates recession processes (or emptying of stores) a little too slowly and recharge processes (filling of stores) a little too quickly. This generates, by comparison with observations, a prolonged interval of low base flow centered on the summer and too short an interval of high base flow centered on the winter. The spectra of discharge from H08 and JULES (but not JULES-TOPMODEL, JULES-PDM, or the other models) also exhibit spectral peaks at the half-annual scale.

The phase spectrum of CLASSIC discharge versus precipitation is very similar to that of observations (Fig. 3h). MPI-HM shows phase values that are too large at the annual scale; that is, the annual cycles in MPI-HM discharge are delayed compared to observations (Fig. 3l), while at higher frequencies the phase spectrum does not follow that of the observations particularly well. LPJmL has a phase spectrum with a similar overall shape to the observations, but between the annual scale and 100 cycles per year (i.e., 3.65-day scale), the phase is less than that observed.

c. *Modeled runoff versus modeled discharge*

The modeled discharge outputs for the Thames basin generally show a high lag-1 autocorrelation similar to that of the observed discharge (Table 1). Figure 4 shows the power spectra of modeled runoff for the Kingston grid box (gray) together with the spectra of modeled discharge (black). The average levels of the runoff spectra are far lower than for the discharge spectra simply because they relate to runoff variability from a single grid box rather than the variability of discharge from the whole basin. The background spectra of modeled runoff are much more similar to those of precipitation than the spectra of modeled discharge. Therefore, the modeling of routing introduces the majority of the increased autocorrelation and attenuation of the subannual variability. The clear exception is provided by GWAVA (Fig. 4) because the discharge time series was created by rescaling the runoff variations (i.e., without routing; section 2). Hence, the lag-1 autocorrelation for GWAVA is correspondingly anomalously low compared to that observed and for other models (Table 1).

An important additional feature of the discharge power spectra in Fig. 4, already noted for LPJmL (Fig. 3m; section 5b), is the presence of a rapid roll off in power at the SCR scale adjacent to the Nyquist frequency, also shown by MATSIRO, JULES, JULES-TOPMODEL, and JULES-PDM. This roll off is not seen in the power spectrum of observed discharge or in the spectra of modeled runoff and is therefore a model

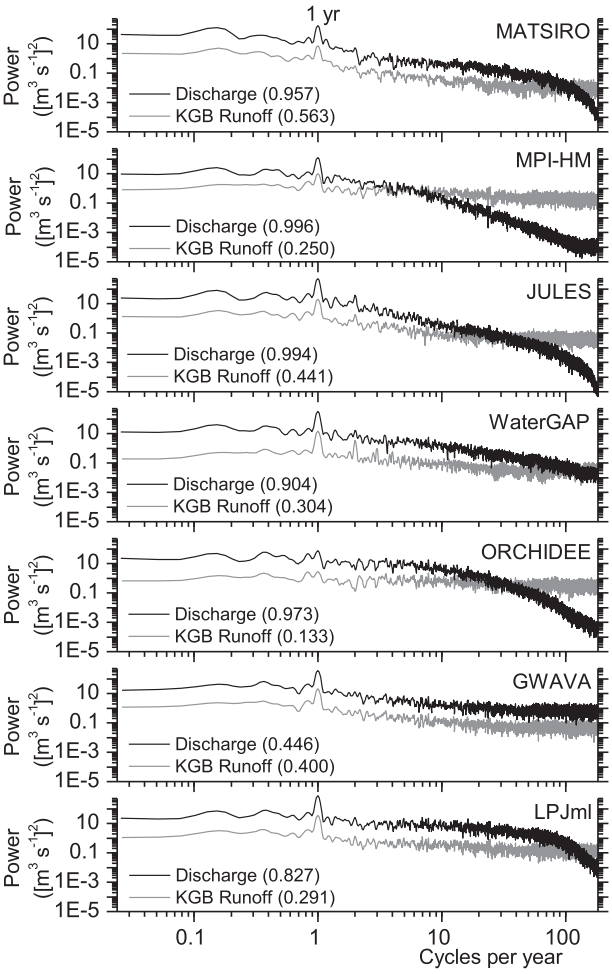

FIG. 4. Power spectra of surface plus subsurface runoff in the Kingston grid box (KGB; gray) and power spectra of modeled discharge (black). The values in parentheses are the lag-1 autocorrelations.

artifact introduced during routing. The roll-off shape of these high-frequency power spectral backgrounds is typical of an AR(1) time series that has been subjected to smoothing via weighted or unweighted averaging of data points that are adjacent in time [e.g., Figs. 3.32 and 5.13 of Weedon (2003)].

d. *Simulating the spectral characteristics of modeled runoff and discharge*

We have simulated the range of spectral characteristics of the model outputs obtained in order to clarify interpretations used later in the evaluations (section 6).

1) POWER SPECTRA

Simple numerical manipulation of the Kingston grid-box precipitation was used to simulate how the models transform the average precipitation variability into discharge variability (appendix B). To allow inspection of

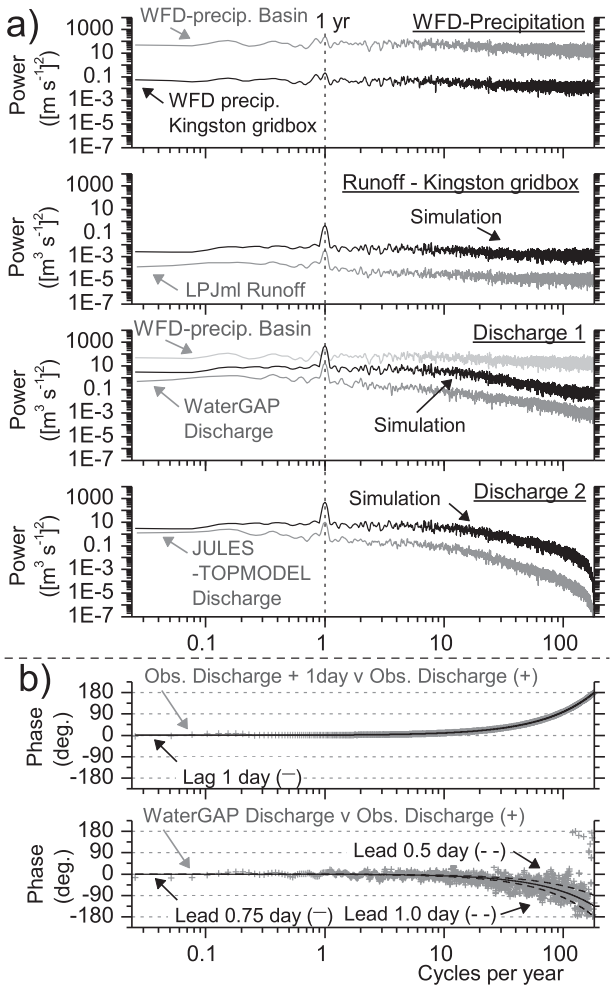

FIG. 5. Simulation of power spectral shapes and phase spectrum trends. (a) Modification of the Kingston gridbox WFD precipitation is used to simulate modeled runoff and discharge. The power spectra of the Kingston WFD precipitation and simulation series are shown in black. Power spectra of model time series outputs are shown in gray but offset vertically from the simulation spectra. Under Discharge 2, the power spectrum of the basinwide WFD precipitation is shown in light gray. (b) Estimated phase spectrum of observed discharge delayed by 1 day vs observed discharge (gray plus symbols). The theoretical phase spectrum due to a lag of 1 day is shown as a black line. Phase spectrum of WaterGAP discharge vs observed discharge is compared to theoretical phase spectra for leads of 0.5 day (upper dashed line), 1 day (lower dashed line), and 0.75 day (full line).

the examples of model power spectra (gray) separately from the simulation spectra (black), the former are offset vertically in Fig. 5a. Runoff was simulated by multiplying the Kingston gridbox precipitation ($m^3 s^{-1}$) by a sinusoid to mimic the effects of the annual cycle in evapotranspiration. This suppresses variability across the spectrum except at the annual scale, producing a power spectrum that looks similar to that of LPJmL runoff for example (Fig. 5a).

Suppression of the high-frequency runoff variability, due to the effects of subsurface storage and water transfer across the basin (sections 5a and 5b; Milly and Wetherald 2002), was simulated by applying a first-order autoregression (appendix B). The specific lag-1 autocorrelation was chosen so that the resulting power spectrum approximately matches that of WaterGAP (Discharge 1; Fig. 5a). The WaterGAP discharge spectrum was illustrated since it is a good match to the observed discharge spectrum (Table 1). Note that, unlike the LPJmL runoff spectrum illustrated in Fig. 5a, most models actually increase autocorrelation substantially during runoff generation, followed by larger increases during routing (Fig. 4).

For all models in Fig. 4, the average levels of the power spectra of runoff from the Kingston grid box alone are far below the average levels for discharge. This simply indicates that routing aggregates basinwide runoff variability at the discharge point. The increased average variability of the discharge compared to gridbox runoff was simulated by multiplying the discharge simulation series by 10.0 (appendix B), thereby approximately reproducing the offsets of runoff and discharge spectra in Fig. 4. This increased level of the simulated discharge spectrum results in greater power than for basinwide WFD precipitation solely at the annual scale (i.e., amplification; Fig. 5a), as noted earlier for the Thames basin observations (section 5a).

Finally, applying a two-point moving average to the simulation series resulted in the roll off in power near the Nyquist frequency exemplified by JULES-TOPMODEL (Discharge 2; Fig. 5a). For JULES, runoff generation was half hourly, but the routing was calculated in daily steps. Potentially, the spurious attenuation of discharge variability at the SCR scale could be alleviated for JULES by using a much shorter time step for the routing calculations (i.e., increasing the Nyquist frequency). However, this solution is not available for models that exhibit the high-frequency roll off but are run entirely at daily steps.

2) PHASE SPECTRA

In theory, modeled routing might generate discharge series that have the wrong phase delays compared to the observed discharge. Offsets of modeled versus observed discharge time series, or delays in discharge variations compared to precipitation variations, produce trends on the phase spectra background described by a simple equation (appendix A; Padilla and Pulido-Bosch 1995). For example, we consider the phase spectrum obtained when the time series of observed discharge is shifted one day later (i.e., a lag of +1) and then compared cross spectrally to the unshifted data. As expected, in Fig. 5b the theoretical trend (black line) using Eq. (A10) passes

TABLE 2. Phase delays at the annual scale. Kingston gridbox modeled runoff vs gridbox WFD precipitation and observed discharge vs observed basinwide precipitation and modeled discharge vs basinwide WFD precipitation; 95% CI = 95% confidence interval.

Observations or model	Kingston gridbox runoff vs Kingston gridbox WFD precipitation (±95% CI)	Discharge vs basinwide precipitation (±95% CI)
Observations	—	+75.6° ± 19.0°
GWAVA	+57.7° ± 13.1°	+67.0° ± 18.3°
LPJmL	+64.7° ± 16.8°	+72.6° ± 20.9°
MATSIRO	+106.2° ± 16.8°	+114.8° ± 21.1°
WaterGAP	+67.9° ± 16.0°	+81.4° ± 19.7°
JULES	+74.7° ± 15.3°	+95.6° ± 22.3°
MPI-HM	+59.5° ± 14.4°	+108.6° ± 23.7°
ORCHIDEE	+7.5° ± 15.7°	+41.6° ± 12.9°

through the phase estimates (gray crosses). Figure 5b also illustrates the phase spectrum of WaterGAP discharge versus observed discharge. The negative trend is well described by fitting a phase shift that corresponds to the modeled discharge leading observed discharge by an average of 0.75 day.

Given this fitting of phase spectrum trends, we infer for the phase spectrum of observed discharge versus observed precipitation (Fig. 3d; section 5a) that the high-frequency trend is at least partly explained by a simple delay. Hence, the Thames basin as observed generates a shift or delay of one day between the discharge output and precipitation input variations as well as phase differences within specific lower-frequency ranges—most obviously at the annual scale. The delay at the annual scale of about 77 days results from substantial delays of the runoff variations caused by the subsurface movement of base flow into the channel.

The subsurface movement of water in the Thames basin is also associated with the increase in autocorrelation (short-term memory) related to attenuation of high-frequency precipitation variability (section 5a). Hence, the observed phase delay at the annual scale is probably linked to the attenuation of variability and associated increased autocorrelation. Modeling of these processes might then be expected to mean that phase delay at the annual scale would occur during both runoff generation (due to within-gridbox subsurface flow) and routing between grid boxes. Table 2 shows the phase at the annual scale of Kingston gridbox runoff compared to the gridbox WFD precipitation as well as the final phase of model discharge compared to the basinwide WFD precipitation. Gridbox runoff data were not available for CLASSIC nor for H08. The table shows that three models (WaterGAP, JULES, and MPI-HM) increase phase delay during both runoff generation and routing as expected. However, allowing for the confidence intervals, three models introduce phase delays solely at the runoff generation stage (MATSIRO and LPJmL plus GWAVA, which did not use routing). This is surprising considering that MATSIRO and LPJmL increased the autocorrelation at both stages (Fig. 4). Conversely, ORCHIDEE only introduces phase delays during routing.

6. Evaluating model performance using modeled versus observed discharge

In this section, we evaluate model performance via amplitude ratio and phase spectra comparing modeled with observed discharge using the inferences from section 5. Modeled discharge is considered to exhibit significant differences to the observations at frequencies where the 95% confidence intervals for amplitude ratios or phase (differences) do not overlap with 1.0 or 0°, respectively. Figure 6 illustrates example cross spectra from the MBE extremes and for JULES and reruns of JULES. Results for all models at the annual scale and averages for the slow response scale, the quick response scale, and the surface runoff and channel routing scale (defined in section 2) are shown in Fig. 7.

Observed daily discharge values in the Thames basin are skewed, with few very high values and many low values. Reanalyzing the data but using Log(modeled discharge) against Log(observed discharge) produced similar biases in amplitude ratios, as shown in Fig. 7 (i.e., mostly overlapping confidence intervals), and virtually identical phase differences. Hence, the skew of the data does not significantly influence the results and interpretations. We also evaluated more variable data from the winter half years (October–March) separately from the less variable summer half years (April–September; Fig. 2). This showed no significant differences in phase biases, and only two models (WaterGAP and H08) have different directions of amplitude bias (i.e., above versus below unity) between the half years.

In the discussion that follows, we seek to identify issues for particular models that could be addressed to improve average performance. However, fully diagnosing the specific causes of the range of issues discussed typically requires detailed knowledge of model structure and parameters.

CLASSIC has slightly too much evaporation, as indicated by the MBE (+11% ± 4%; Table 1), and the overall variability, as indicated by the standard deviation ($\sigma = 73\,m^3\,s^{-1}$), is relatively close to observations ($\sigma = 67\,m^3\,s^{-1}$; Table 1). It has very good agreement with observations in terms of amplitude and phase at the SR, QR, and SCR scales with slightly too much variation at the annual scale, though good timing. There is a slight deviation from sinusoidal variations of discharge at the

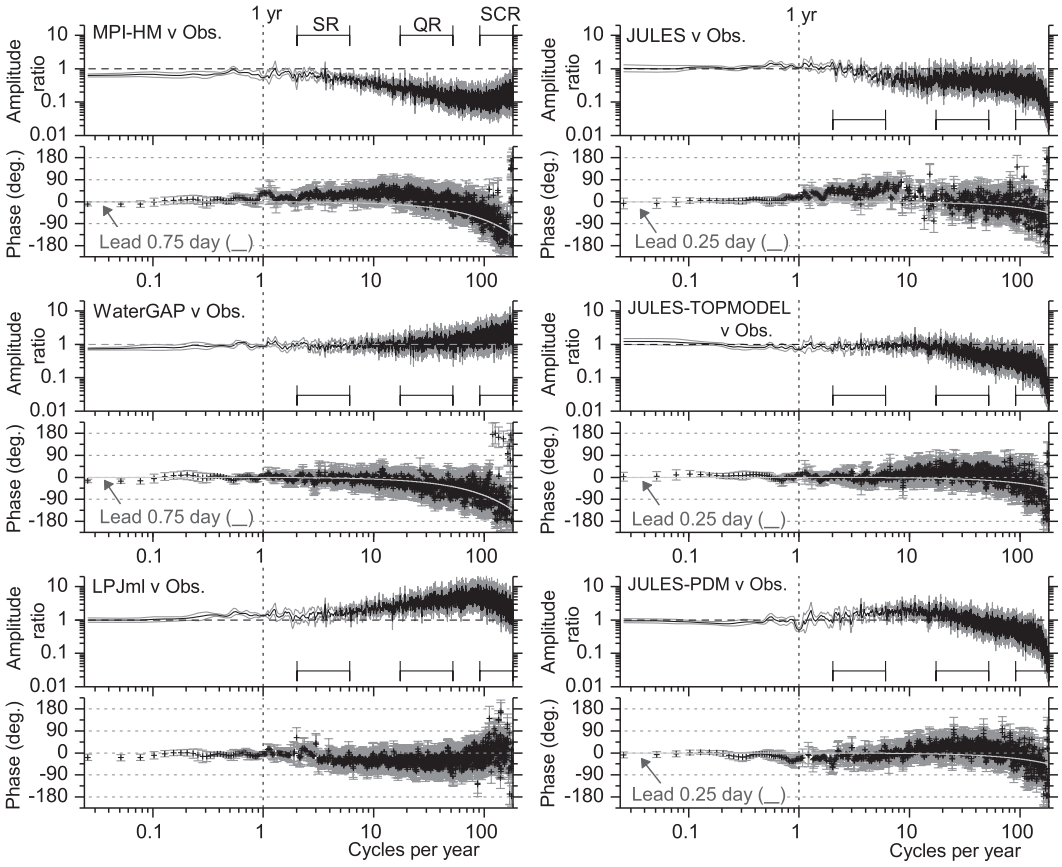

FIG. 6. Cross-spectral results (amplitude-ratio spectra and phase spectra) comparing modeled with observed daily naturalized discharge for the Thames basin (1963–2001). The 95% confidence intervals are indicated in gray. Horizontal bars indicate named frequency bands (with period ranges): SR is 182.5–60 days, QR is 21–7 days, and SCR is 4–2 days. Theoretical phase spectra for offsets (leads) of the modeled vs observed discharge time series are shown for selected cases using light gray lines.

annual scale (long, low base flow intervals and short, high base flow intervals) generating a spectral peak at the half-annual scale (Fig. 3e). The good overall performance in cross-spectral terms is expected given that the model is catchment based and calibrated (though not calibrated with the WFD).

WaterGAP is also calibrated so the water balance is closed (MBE = −4% ± 3%), suggesting an accurate estimation of the amount of evaporation and an average variability across all scales similar to that for the observations (σ = 63 $m^3 s^{-1}$). There is slightly too little variation at the annual scale (amplitude ratio = 0.88 ± 0.02), good agreement with observations at the SR and QR scales, and too much variation at the SCR scale with associated early phase. Fitting the subannual part of the phase spectrum indicates that the modeled discharge arrives, on average, about three-quarters of a day earlier than observed discharge (Figs. 4, 5). These average results mask the fact, revealed from analyzing the half years of data separately, that at the QR scale there is too much variability in the winter (i.e., high base flow) half years and too little variability in the summer half years (cf. Fig. 2). Hence, the direction of model bias in the amplitude ratio for WaterGAP depends on the average flow conditions, as observed for some lumped models (Herman et al. 2013). This demonstrates that finding amplitude ratios close to 1.0 and phase differences indistinguishable from 0° when studying the whole time series does not guarantee correct model behavior.

Although H08 has very good average evaporation (MBE = +1% ± 8%), it shows far too much variability in discharge (σ = 127 $m^3 s^{-1}$). The annual scale amplitude is far too large (amplitude ratio = 1.94 ± 0.08) and too early (phase = −23.6° ± 3.2°), and there is a very pronounced harmonic spectral peak at the half-annual scale, denoting strongly nonsinusoidal annual-scale variations (too long intervals of low base flow, too short intervals of high base flow). At the SR and QR scales, the phase is reasonable, but again the amplitude ratio is above 1.0 (Fig. 7). At the SCR scale, both the

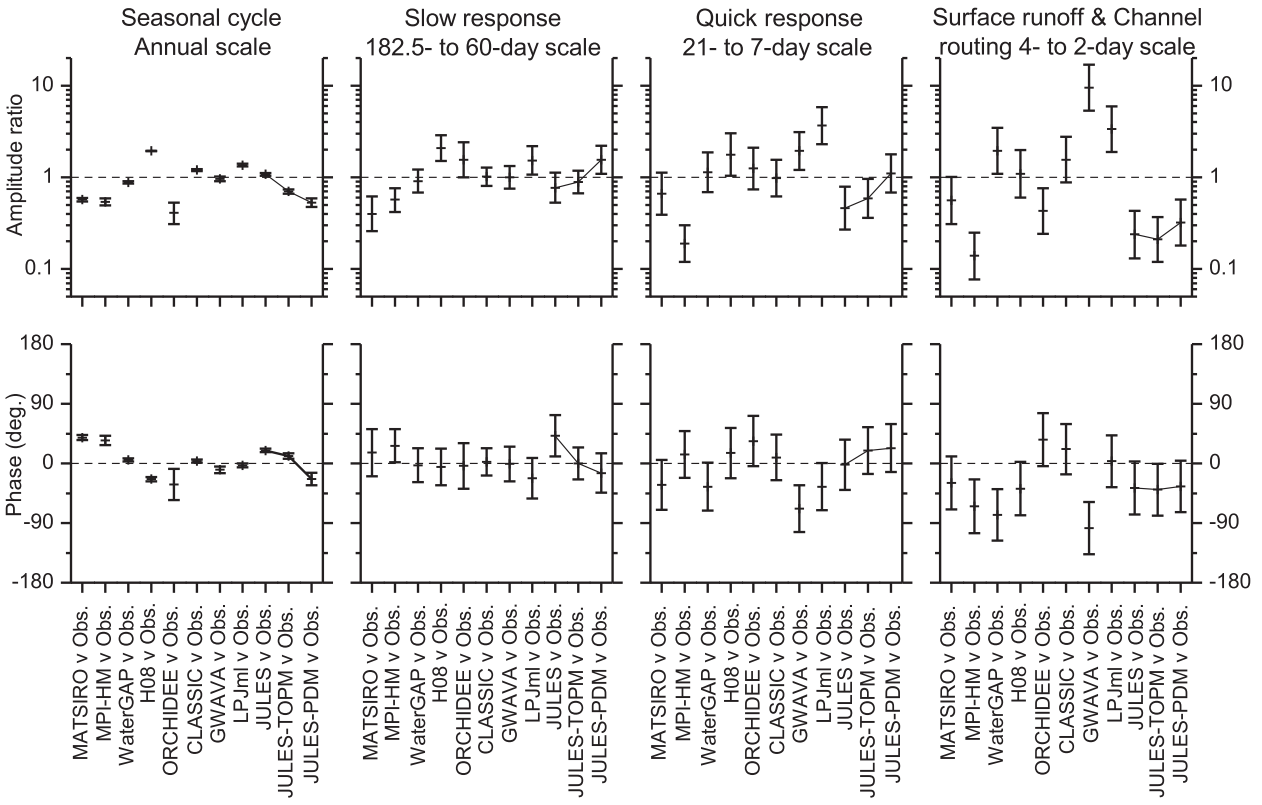

FIG. 7. Cross-spectral results from comparing modeled with observed discharge at the annual scale and averaged over different frequency bands. Negative phase indicates that, on average, the modeled discharge variations are too early compared to observed discharge variations. Vertical bars denote the 95% confidence intervals for the mean amplitude ratio and for the mean phase difference in each case. Within the frequency bands, the average phase is calculated at frequencies where the coherency exceeds the 95% significance level.

variability and phase are reasonable. As for WaterGAP, at the QR scale the amplitude is too large in the winter half year, but too small in the summer half year. Unlike WaterGAP, this is also true at the SR and annual scales.

There is far too little evaporation in LPJmL (MBE = +28% ± 5%) with too much variability ($\sigma = 128\,m^3\,s^{-1}$). Amplitudes are too high from the annual to the SCR scale, although the phase is generally reasonable. Since the excess amplitude is found across the annual and higher-frequency part of the spectrum, it is reasonable to infer for this model that the excess variability is due to the lack of evaporation (determining MBE) combined with too little subsurface storage. The lack of subsurface storage limits the amount of modeled high-frequency attenuation leading to too little autocorrelation. The delay in phase at the annual scale is entirely introduced during runoff generation rather than, as expected, partly during routing [section 5d(2), Table 2]. The roll off in power at the SCR scale (not seen in the spectrum of observed discharge; Fig. 3a) indicates attenuation due to a moving-average process introduced during routing [section 5d(1)].

GWAVA also has too little evaporation (MBE = +22% ± 5%) and too much variability ($\sigma = 94\,m^3\,s^{-1}$). However, the amplitude and phase of GWAVA discharge are among the best modeled at the annual and SR scales (Fig. 7). The excessive variability and early phase at the QR and SCR scales are readily explained by the lack of routing (section 2; Fig. 4).

MPI-HM has too much evaporation (MBE = −31% ± 3%) and too little variability ($\sigma = 37\,m^3\,s^{-1}$). The low amplitudes compared to observations are seen from the annual to the SCR scale with phase ranging from too late at the annual scale to too early at the SCR scale. The lack of variability across the spectrum is apparently explicable simply as due to the excessive evaporation. The late (positive) phases at the annual to SR scales might be associated with modeled response times that are too long within the subsurface stores.

MATSIRO also has too much evaporation (MBE = −33% ± 7%) and too little average variability ($\sigma = 50\,m^3\,s^{-1}$). However, the amplitude ratio is well below unity only at the annual and SR scales. The phase is far too late at the annual scale, but otherwise consistent (within error) with the observations at higher frequencies. The phase delay at the annual scale is entirely introduced at the runoff generation stage (Table 2), and routing includes

spurious attenuation at the SCR scale in association with a moving average process [Fig. 4, section 5d(1)].

The overall performance of JULES depends on the configuration used (section 3). JULES-TOPMODEL and JULES exhibit too much evaporation (MBE = −27% ± 4% and −17% ± 4%, respectively) and less, and slightly less, average variance than the observations (σ = 56 and 64 $m^3 s^{-1}$), while JULES-PDM shows opposite characteristics (MBE = +9% ± 5%, σ = 76 $m^3 s^{-1}$). All configurations have the roll off in power at the SCR scale associated with excessive high-frequency attenuation due to averaging the discharge values from successive time steps during the routing calculations. This processing probably also explains the high-frequency trend in phase toward the SCR scale that can be modeled as due to discharge variations being an average of 0.25 days early compared to observations (Fig. 6).

Discharge from JULES has amplitude variations that are consistent with observations at the annual and SR scales but with phase that is too late (positive). Slightly nonsinusoidal annual cycles produce a harmonic spectral peak at the half-annual scale (low-flow base flow intervals too long, high base flow intervals too short). There is too little variability at the QR scale, but the phase is reasonable. The late phase at the SR and annual scales may indicate residence times that are too long in the slow subsurface stores.

When TOPMODEL is implemented, a proportion of the precipitation is retained in the surface soil stores that would otherwise have been transferred into the subsurface during the JULES run. This means that from the SR to QR scales, the variability of discharge from JULES-TOPMODEL agrees with the observations better than the JULES run. The phases at the annual and SR scales are also improved compared to observations and the JULES output. However, the extra water available in the soil stores allows more evaporation and transpiration, so the MBE becomes more negative than for JULES. At the annual scale, the variability is too small (unlike the JULES run), probably because of the extra water evaporated over subannual time scales.

When PDM is implemented in JULES, more water is diverted to surface runoff, reducing the overall evaporative losses from the (shallow) soil stores, so the MBE is positive (rather than negative for JULES and JULES-TOPMODEL). At the QR scale, the amplitude and phase agree within error with the observations, an improvement compared to JULES. At the SR scale, the discharge variations are too large, though of the right phase. Additionally, the excess variability at the SR scale means that annual-scale variations in discharge are too small and occur too early, representing a worse result at this scale compared to the JULES run. Note that JULES-PDM was run using default values of 1.0 m for the soil depth parameter and b = 1.0 for the shape parameter (Moore and Clarke 1981): there was no attempt to improve the results by calibrating the parameters to suit the Thames basin (cf. Clark and Gedney 2008).

ORCHIDEE has a similar performance to JULES-PDM (MBE = +7% ± 5%, σ = 77 $m^3 s^{-1}$). At the annual SR and QR scales, the cross-spectral results are very similar to JULES-PDM, and indeed the time series appear very similar (Fig. 2). However, ORCHIDEE does not introduce the excessive high-frequency amplitude suppression during routing seen for JULES (and MATSIRO and LPJmL). On the other hand, at the annual scale, the phase delay, which is too small (phase negative), is introduced entirely during routing (Table 2).

7. Conclusions

The simulation of discharge rates in the distributed models applied to the Thames basin requires accurate modeling of evaporative losses that can be assessed using MBE. We have demonstrated that the cross-spectral methods used are appropriate for assessing the relative variability and timing of modeled versus observed discharge. MBE needs to be assessed alongside the cross-spectral results because the overall water balance cannot be determined by the spectral methods because of the subtraction of the mean from each time series during preprocessing (i.e., linear detrending; appendix A). Note that observing amplitude ratios close to 1.0 and phase differences indistinguishable from 0° when studying the whole time series does not guarantee correct model behavior (see WaterGAP results in section 6). Nevertheless, significant deviations from these reference levels can be used to focus attention on problems with the representation of specific physical processes by a model.

The evaluations of model performance in section 6, based on amplitude ratio and phase spectra comparing modeled with observed discharge, was predicated on the simulations of model spectral properties in section 5. Rerunning JULES using TOPMODEL or PDM for improving subgrid heterogeneity does help with the amplitude and phase of discharge at between half-year to 7-day periods. However, the cross spectra show how these reconfigurations compromise the otherwise good amplitude performance at the annual scale.

The evaluations showed that, in addition to the effects on discharge variability of too little or too much evaporation, the capacity of surface and/or subsurface stores and time constants are not appropriate in some models. As well as the need for implementation of daily routing for GWAVA, specific model performance could be improved by also tackling a variety of issues. These

issues are 1) nonsinusoidal annual cycles of discharge (CLASSIC, JULES, and H08), 2) excessive attenuation of highest-frequency variability (MATSIRO, LPJmL, and JULES in all three configurations), 3) excessive variability in discharge during winter half years but too little variability in summer half years (WaterGAP and H08), and 4) introduction of annual phase delays only during runoff generation (MATSIRO, GWAVA, and LPJmL) or only during routing (ORCHIDEE).

Acknowledgments. We thank two anonymous referees for their constructive comments. Doug Clark (CEH) helped in discussions of routing and discharge generation within WATCH/WaterMIP and JULES. G.P.W. and M.J.B. were supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). C.P. and S.C. were supported by the NERC-CEH National Capability funding (Water and Pollution Science Management Group).

APPENDIX A

Spectral and Cross-Spectral Estimation

To avoid power leakage from the zero-frequency component (Percival and Walden 1993, 504–506), all time series were initially detrended linearly, removing any trend in the mean and leaving the data mean centered. Split cosine tapering of the first and last 10% of the centered data was used to suppress periodogram leakage (Priestley 1981; von Storch and Zwiers 1999). Weedon (2003) compares methods of spectral estimation and provides sources of algorithms for standard time series methods in the appendix.

For a mean-centered, tapered, finite time series $X(t)$ consisting of N values at discrete time steps t with a fixed time-step interval of Δt, the power spectrum is evaluated at the discrete Fourier frequencies f defined in terms of the proportion of the full dataset length (Percival and Walden 1993) with $f = i/N$. The integer i provides the frequency index or harmonic number and runs from 0 to $N/2$. For plotting results, the absolute frequency F is related to the Fourier frequency using $F = f/\Delta t$. In this case, for values at daily time steps and absolute frequency expressed in units of cycles per year, the sample interval Δt, allowing for leap days, is 1.0/365.24.

A mean-centered time series can be represented (Percival and Walden 1993) in terms of the Fourier frequencies as

$$X(t) = \sum_{f=1/N}^{1/2} [A(f)\cos(2\pi ft) + B(f)\sin(2\pi ft)], \qquad \text{(A1)}$$

for $t = 1, 2, 3, \ldots, N$, where $A(f)$ is the cosine amplitude and $B(f)$ is the sine amplitude. Note that here the time-step and frequency indices indicate discrete sequences, not continuous functions.

In the periodogram approach, the cosine and sine amplitudes are estimated (Ifeachor and Jervis 1993; Percival and Walden 1993) using

$$A(f) = \frac{2}{N}\sum_{t=1}^{N}[X(t)\cos(2\pi ft)] \qquad \text{(A2)}$$

and

$$B(f) = \frac{2}{N}\sum_{t=1}^{N}[X(t)\sin(2\pi ft)]. \qquad \text{(A3)}$$

The periodogram power estimates $I(f)$ indicating the power or variance at the Fourier frequencies are obtained using

$$I(f) = A(f)^2 + B(f)^2. \qquad \text{(A4)}$$

To allow analysis of data from winter half years separately from summer half years (i.e., nonuniform time steps between data points; section 6), we used the Lomb–Scargle periodogram spectral estimates from the program PERIOD of Press et al. (1992). Periodogram estimates, with just 2 degrees of freedom, are distributed erratically around any theoretical spectral background noise level. The Tukey–Hanning spectral window (Priestley 1981) was applied three times to the periodogram to yield power spectral estimates $Gxx(f)$ with 8 degrees of freedom.

One-sided power spectral confidence levels were obtained using a standard chi-squared distribution allowing for the degrees of freedom (Priestley 1981; Percival and Walden 1993). Quasi-periodic components, especially at the scale of the annual cycle, were identified as power-spectral peaks emerging above the 99.0% and the 99.999% confidence levels relative to the locally defined spectral background (estimated via moving window averaging; Press et al. 1992). The higher confidence level quoted corresponds to the false alarm probability α' (by applying the Šidák correction to the target probability level α—this is analogous to a Bonferroni correction for multiple tests; Abdi 2007).

The first step in generation of the coherency spectrum for comparing two time series [$X(t)$ and $Y(t)$] is estimation of the coperiodogram $C^Ixy(f)$ and quadrature periodogram $Q^Ixy(f)$ (Priestley 1981) via

$$C^Ixy(f) = Ax(f)Ay(f) + Bx(f)By(f) \qquad \text{(A5)}$$

and

$$Q^Ixy(f) = Bx(f)Av(f) - Ax(f)By(f). \qquad \text{(A6)}$$

The coperiodogram and quadrature periodogram were smoothed using a Tukey–Hanning spectral window as for the power spectra, producing the estimated cospectrum $Cxy(f)$ and estimated quadratic spectrum $Qxy(f)$. The estimated cross-amplitude spectrum $Gxy(f)$ is obtained using (Priestley 1981; von Storch and Zwiers 1999)

$$Gxy(f) = [Cxy(f)^2 + Qxy(f)^2]^{1/2}. \quad \text{(A7)}$$

The estimated coherency spectrum $\mathrm{Coh}(f)$ is then derived (Priestley 1981) as

$$\mathrm{Coh}(f) = \frac{Gxy(f)}{[Gxx(f)Gyy(f)]^{1/2}}. \quad \text{(A8)}$$

The phase spectrum $\Phi(f)$ indicates, for each frequency, the relative difference in timing of oscillations in paired time series. In terms of radians, it is derived as

$$\Phi(f) = \tan2^{-1}[-Qxy(f)/Cxy(f)]. \quad \text{(A9)}$$

The inverse arctangent (or computational function ATAN2) limits the phase differences to between $+\pi$ and $-\pi$ radians. Phase (difference) in degrees equals $\Phi(f) \times 360.0/2\pi$ radians, so the central estimates are restricted to between $-180°$ and $+180°$. The 95% confidence interval for phase expands rapidly at low coherency (von Storch and Zwiers 1999), so phase is only illustrated and used in the frequency band averages of Fig. 7, where coherency exceeds the 95% coherency significance level (0.78 here). This limits the phase uncertainty for the central estimates plotted to $\leq \pm 51°$.

As discussed in the text [section 5d(2)], we explore the phase shift $\Delta\Phi(f)$ (radians) due to an offset, positive or negative between a time series and itself λ (from $-N$ to N in time-step units). The phase shift (Priestley 1981) simply depends on the frequency index i ($=fN$) multiplied by the proportion of the full time series represented by the offset (i.e., λ/N):

$$\Delta\Phi(f) = \tan2^{-1}[\sin(fN2\pi\lambda/N)/\cos(fN2\pi\lambda/N)]. \quad \text{(A10)}$$

The power at each frequency from windowed periodogram estimates is derived from squared amplitude values (note that for some estimation methods the power is given by the area under the power spectrum; Priestley 1981). Hence, the amplitude-ratio spectra $R(f)$ were derived as

$$R(f) = Gxx(f)^{1/2}/Gyy(f)^{1/2}. \quad \text{(A11)}$$

In comparing the amplitudes of input time series and output time series via the spectral transfer function, the gain spectrum $\mathrm{Gain}(f)$ is obtained [section 9.2 of Priestley (1981)] using

$$\mathrm{Gain}(f) = Gxy(f)/Gyy(f). \quad \text{(A12)}$$

The gain at each frequency expresses the change in amplitude of the output from a system (e.g., discharge) compared to the amplitude of the input (e.g., precipitation). Multiplying the right-hand sides of Eqs. (A8) and (A11) yields

$$R(f)\mathrm{Coh}(f) = \mathrm{Gain}(f). \quad \text{(A13)}$$

APPENDIX B

Simulating Modeled Runoff and Discharge Power Spectral Characteristics

The WFD Kingston gridbox precipitation $\mathrm{PN}_{\mathrm{King}}(t)$ was rescaled from millimeters per day to cubic meters per second. A cosine wave time series with an absolute frequency of 1 cycle per year was generated, multiplied by 0.25, and then +0.25 was added to all time steps. The resulting series ranges from +0.5 in midwinter to 0.0 in midsummer. Multiplication of the rescaled precipitation data by the cosine series (imposed amplitude modulation) partly suppresses midwinter precipitation but severely attenuates midsummer precipitation variations, yielding surface plus subsurface runoff:

$$\mathrm{RO}(t) = \mathrm{PN}_{\mathrm{King}}(t) \times \{0.25 + [0.25\cos(2\pi f_1 t)]\}. \quad \text{(B1)}$$

The Fourier frequency f_1 within Eq. (B1) is calculated via the absolute frequency F (=1.0 cycle per year) and sample rate Δt:

$$f_1 = \Delta t F = 1.0(1.0/365.24). \quad \text{(B2)}$$

Discharge was simulated by imposing first-order autoregression to mimic the attenuation of high-frequency variations due to subsurface storage and transport across the basin during routing. The lag-1 autocorrelation ($\rho_1 = 0.7$) was selected so that the power spectrum of the simulated discharge provided a reasonable match to that of WaterGAP—itself a good match to the spectrum of observed discharge. Routing of the runoff to the discharge point increases average variability (Fig. 4). Increased variability due to routing was simulated by multiplying the autoregressed series by 10.0 to approximate the offset in levels of the runoff and discharge spectra in Fig. 4:

$$Q(t) = 10.0\{\mathrm{RO}(t) + [0.7\mathrm{RO}(t-1)]\}. \quad \text{(B3)}$$

Some models analyzed apparently partially represent routing by accumulating water from adjacent grid boxes

from different time steps. The effect of such processing on the simulated discharge was applied using a simple (unweighted or boxcar) moving average:

$$Q'(t) = [Q(t) + Q(t-1)]/2. \quad \text{(B4)}$$

REFERENCES

Abdi, H., 2007: The Bonferroni and Šidák corrections for multiple comparisons. *Encyclopedia of Measurement and Statistics,* N. Salkind, Ed., Sage, 103–107.

Best, M. J., and Coauthors, 2011: The Joint UK Land Environment Simulator (JULES) model description—Part 1: Energy and water fluxes. *Geosci. Model Dev.,* **4,** 677–699, doi:10.5194/gmd-4-677-2011.

Beven, K. J., M. J. Kirkby, N. Schofield, and A. F. Tagg, 1984: Testing a physically-based flood forecasting model (TOPMODEL) for three UK catchments. *J. Hydrol.,* **69,** 119–143, doi:10.1016/0022-1694(84)90159-8.

Billings, S. A., 2013: *Nonlinear System Identification: NARMAX methods in the Time, Frequency, and Spatio–Temporal Domains.* Wiley, 555 pp.

Bryce, R. M., and K. B. Sprague, 2012: Revisiting detrended fluctuation analysis. *Sci. Rep.,* **2,** 315, doi:10.1038/srep00315.

Clark, D. B., and N. Gedney, 2008: Representing the effects of subgrid variability of soil moisture on runoff generation in a land surface model. *J. Geophys. Res.,* **113,** D10111, doi:10.1029/2007JD008940.

Crooks, S. M., and P. S. Naden, 2007: CLASSIC: A semi-distributed rainfall–runoff modelling system. *Hydrol. Earth Syst. Sci.,* **11,** 516–531, doi:10.5194/hess-11-516-2007.

Döll, P., and B. Lehner, 2002: Validation of a new global 30-min drainage direction map. *J. Hydrol.,* **258,** 214–231, doi:10.1016/S0022-1694(01)00565-0.

Ebisuzaki, W., 1997: A method to estimate the statistical significance of a correlation when data are serially correlated. *J. Climate,* **10,** 2147–2153, doi:10.1175/1520-0442(1997)010<2147:AMTETS>2.0.CO;2.

Fleming, S. W., 2014: A non-uniqueness problem in the identification of power-law spectral scaling for hydroclimatic time series. *Hydrol. Sci. J.,* **59,** 73–84, doi:10.1080/02626667.2013.851384.

Gedney, N., and P. M. Cox, 2003: The sensitivity of global climate model simulations to the representation of soil moisture heterogeneity. *J. Hydrometeor.,* **4,** 1265–1275, doi:10.1175/1525-7541(2003)004<1265:TSOGCM>2.0.CO;2.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.,* **113,** D06104, doi:10.1029/2007JD008972.

Gudmundsson, L., and Coauthors, 2012a: Comparing large-scale hydrological model simulations to observed runoff percentiles in Europe. *J. Hydrometeor.,* **13,** 604–620, doi:10.1175/JHM-D-11-083.1.

——, T. Wagner, L. M. Tallaksen, and K. Engeland, 2012b: Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe. *Water Resour. Res.,* **48,** W11504, doi:10.1029/2011WR010911.

Haddeland, I., and Coauthors, 2011: Multimodel estimate of global terrestrial water balance: Setup and first results. *J. Hydrometeor.,* **12,** 869–884, doi:10.1175/2011JHM1324.1.

Hanasaki, N., S. Kanae, T. Oki, K. Masuda, K. Motoya, N. Shirakawa, Y. Shen, and K. Tanaka, 2008: An integrated model for the assessment of global water resources—Part 1: Model description and input meteorological forcing. *Hydrol. Earth Syst. Sci.,* **12,** 1007–1025, doi:10.5194/hess-12-1007-2008.

Heneghan, C., and G. McDarby, 2000: Establishing the relation between detrended fluctuation analysis and power spectral density analysis for stochastic processes. *Phys. Rev. E,* **62,** 6103–6110.

Herman, J. D., P. M. Reed, and T. Wagener, 2013: Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior. *Water Resour. Res.,* **49,** 1400–1414, doi:10.1002/wrcr.20124.

Hoskins, J. R. M., 1984: Modeling persistence in hydrological time series using fractional differencing. *Water Resour. Res.,* **20,** 1898–1908, doi:10.1029/WR020i012p01898.

Hurst, H. E., 1951: Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civ. Eng.,* **116,** 770–808.

Ifeachor, E. C., and B. W. Jervis, 1993: *Digital Signal Processing: A Practical Approach.* Addison-Wesley, 760 pp.

Jenkins, G. M., and D. G. Watts, 1969: *Spectral Analysis and Its Applications.* Holden Day, 525 pp.

Jones, S. B., 1983: The estimation of catchment average point rainfall profiles. IH Rep. 87, Institute of Hydrology, Wallingford, United Kingdom, 34 pp. [Available online at www.ceh.ac.uk/products/publications/documents/ihreportno87lo-res.pdf.]

Kantelhardt, J. W., E. Koscielny-Bunde, D. Ryski, P. Braun, A. Bunde, and S. Havlin, 2006: Long-term persistence and multifractality of precipitation and river runoff records. *J. Geophys. Res.,* **111,** D01106, doi:10.1029/2005JD005881.

Labat, D., 2010: Cross wavelet analyses of annual continental freshwater discharge and selected climate indices. *J. Hydrol.,* **385,** 269–278, doi:10.1016/j.jhydrol.2010.02.029.

——, R. Ababou, and A. Mangin, 2000a: Rainfall–runoff relations for karstic springs. Part I: Convolution and spectral analysis. *J. Hydrol.,* **238,** 123–148, doi:10.1016/S0022-1694(00)00321-8.

——, ——, and ——, 2000b: Rainfall–runoff relations for karstic springs. Part II: Continuous wavelet transform and multiresolution analyses. *J. Hydrol.,* **238,** 149–178, doi:10.1016/S0022-1694(00)00322-X.

Lane, S. N., 2007: Assessment of rainfall–runoff models based upon wavelet analysis. *Hydrol. Processes,* **21,** 586–607, doi:10.1002/hyp.6249.

Liu, Y., J. Brown, J. Demargne, and D.-J. Seo, 2011: A wavelet-based approach to assessing timing errors in hydrologic predictions. *J. Hydrol.,* **397,** 210–224, doi:10.1016/j.jhydrol.2010.11.040.

Marsh, T. J., and J. Hannaford, 2008: *UK Hydrometric Register.* Hydrological Data UK Series, Centre for Ecology and Hydrology, 210 pp.

Materia, S., P. A. Dirmeyer, Z. Guo, A. Alessandri, and A. Navarra, 2010: The sensitivity of simulated river discharge to land surface representation and meteorological forcings. *J. Hydrometeor.,* **11,** 334–351, doi:10.1175/2009JHM1162.1.

Mesa, O. J., and G. Poveda, 1993: The Hurst effect: The scale fluctuation approach. *Water Resour. Res.,* **29,** 3995–4002, doi:10.1029/93WR01686.

Milly, P. C. D., and R. T. Wetherald, 2002: Macroscale water fluxes 3. Effects of land processes on variability of monthly river discharge. *Water Resour. Res.,* **38,** 1235, doi:10.1029/2001WR000761.

Montanari, A., and E. Toth, 2007: Calibration of hydrological models in the spectral domain: An opportunity for scarcely gauged basins? *Water Resour. Res.,* **43,** W05434, doi:10.1029/2006WR005184.

Moore, R. J., and R. T. Clarke, 1981: A distribution function approach to rainfall runoff modelling. *Water Resour. Res.,* **17,** 1367–1382, doi:10.1029/WR017i005p01367.

Moussu, F., L. Oudin, V. Plagnes, A. Mangin, and H. Bedjoudi, 2011: A multi-objective calibration framework for rainfall–discharge models applied to karst systems. *J. Hydrol.,* **400,** 364–376, doi:10.1016/j.jhydrol.2011.01.047.

Mudelsee, M., 2007: Long memory of rivers from spatial aggregation. *Water Resour. Res.,* **43,** W01202, doi:10.1029/2006WR005721.

Naden, P. S., 1992: Spatial variability in flood estimation for large catchments: The exploitation of channel network structure. *Hydrol. Sci. J.,* **37,** 53–71, doi:10.1080/02626669209492561.

Nash, J. E., and J. V. Sutcliffe, 1970: River flow forecasting through conceptual models part I—A discussion of principles. *J. Hydrol.,* **10,** 282–290, doi:10.1016/0022-1694(70)90255-6.

Padilla, A., and A. Pulido-Bosch, 1995: Study of hydrographs of karstic aquifers by means of correlation and cross-spectral analysis. *J. Hydrol.,* **168,** 73–89, doi:10.1016/0022-1694(94)02648-U.

Pauwels, V. R. N., and J. M. De Lannoy, 2011: Multivariate calibration of a water and energy balance model in the spectral domain. *Water Resour. Res.,* **47,** W07523, doi:10.1029/2010WR010292.

Pelletier, J. D., and D. L. Turcotte, 1997: Long-range persistence in climatological and hydrological time series: Analysis, modelling and application to drought hazard assessment. *J. Hydrol.,* **203,** 198–208, doi:10.1016/S0022-1694(97)00102-9.

Percival, D. B., and A. T. Walden, 1993: *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques.* Cambridge University Press, 583 pp.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in Fortran: The Art of Scientific Computing.* Cambridge University Press, 963 pp.

Priestley, M. B., 1981: *Spectral Analysis and Time Series.* Academic Press, 890 pp.

Quets, J. J., G. J. M. De Lannoy, and V. R. N. Pauwels, 2010: Comparison of spectral and time domain calibration methods for precipitation–discharge processes. *Hydrol. Processes,* **24,** 1048–1062, doi:10.1002/hyp.7546.

Schaefli, B., and E. Zehe, 2009: Hydrological model performance and parameter estimation in the wavelet-domain. *Hydrol. Earth Syst. Sci.,* **13,** 1921–1936, doi:10.5194/hess-13-1921-2009.

Schepers, H. E., J. H. G. M. van Beek, and J. B. Bassingthwaighte, 1992: Four methods to estimate the fractal dimension of self-affine signals. *IEEE Eng. Med. Biol. Mag.,* **11,** 57–64, doi:10.1109/51.139038.

Smith, L. C., D. L. Turcotte, and B. L. Isacks, 1998: Stream flow characterization and feature detection using a discrete wavelet transform. *Hydrol. Processes,* **12,** 233–249, doi:10.1002/(SICI)1099-1085(199802)12:2<233::AID-HYP573>3.0.CO;2-3.

Taylor, K. E., 2001: Summarizing multiple model performance in a single diagram. *J. Geophys. Res.,* **106,** 7183–7192, doi:10.1029/2000JD900719.

von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research.* Cambridge University Press, 484 pp.

Weedon, G. P., 2003: *Time-Series Analysis and Cyclostratigraphy: Examining Stratigraphic Records of Environmental Cycles.* Cambridge University Press, 259 pp.

——, and Coauthors, 2011: Creation of the WATCH forcing data and its use to assess global and regional reference crop evaporation over land during the twentieth century. *J. Hydrometeor.,* **12,** 823–848, doi:10.1175/2011JHM1369.1.

Willmott, C. J., and K. Matsuura, 2005: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Res.,* **30,** 79–82, doi:10.3354/cr030079.