Supplementary Material for

A Multiple Imputation Strategy for Eddy-Covariance Data

Domenico Vitale, Massimo Bilancia, Dario Papale

Appendix A: Regime detection procedure

As highlighted by Doove et al. (2014)[1], caution is needed when data contain nonlinear structures like interaction terms between climate and hydrological variables, that are omitted from the imputation model and are likely to create serious biases in imputed values. Therefore, in order to take full advantage of the MVN modelling and alleviate the effect of nonlinearity and interactions, we considered a data-driven procedure to partition a calendar year into a series of stable regimes, each one showing limited variations both in climate variables and hydrological components. For this purpose, most studies involving EC data employ calendar months, making use of the climatic definition of seasons with a fixed start and end date (winter: JFM; spring: AMJ; summer, JAS; autumn: OND) or adopt rolling and overlapping temporal windows of 15/30 days length. While the first of the two approaches does not guarantee the ecological regimes are correctly detected, the latter could be inefficient from a computational point of view.

As models should preferably be fitted to the data in an automatic fashion without unnecessary user involvement, we have chosen to make use of a two-step change-point detection procedure consisting of: (i) a smoothing-spline regression on NEE daytime subset to extract the long-period

[1] Doove, L.L., Van Buuren, S., Dusseldorp, E. (2014). Recursive Partitioning for Missing Data Imputation in the Presence of Interaction Effects. Computational Statistics & Data Analysis. 72, 92–104. doi:10.1016/j.csda.2013.10.025

component; (ii) a breakpoint analysis to detect shifts in mean level of the NEE fitted values (Zeileis et al., 2002)[2]. The workflow of this procedure is represented in Figure A1.

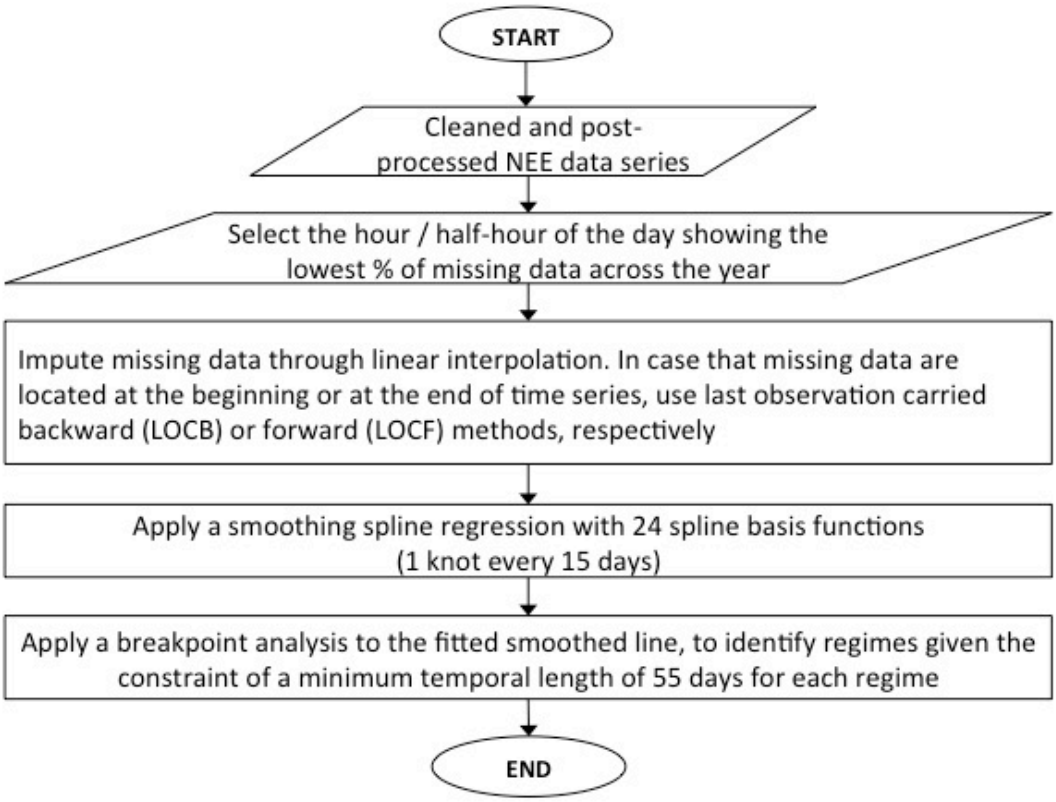

Figure A1. Work flow of the regime detection procedure for EC datasets.

It is important to stress that the proposed procedure is mainly intended to split EC datasets into more homogeneous periods, in order (i) to preserve the short-term relationship between EC fluxes and climate variables (Hui et al., 2004)[3] and (ii) to assure a sufficient number of data points inside each

[2] Zeileis, A., Leisch, F., Hornik, K., Kleiber, C. (2002). Strucchange : An R Package for Testing for Structural Change in Linear Regression Models. Journal of Statistical Software. 7(2). doi:10.18637/jss.v007.i02

[3] Hui, D., Wan, S., Su, B., Katul, G., Monson, R., Luo, Y. (2004). Gap-Filling Missing Data in Eddy Covariance Measurements Using Multiple Imputation (MI) for Annual Estimations. Agricultural and Forest Meteorology. 121(1-2), 93–111. doi:10.1016/S0168-1923(03)00158-8

regime, enough to run MI algorithms without instability due to lack of information. Although our intent is not to identify regimes with a strongly characterized relevance from an ecological point of view, we found that in most cases our results were in good agreement with the hydro-ecological regime definitions proposed by Thomas et al. (2009)[4].

As an example, the following six regimes were identified for the IT-CA1 use case (Figure A2):

- Regime 1 – Winter/dormant period, starting from DoY 1 up to DoY 58.
- Regime 2 – Growing season with non-limited water resources, from DoY 59 to DoY 112. This phase corresponds to poplar leaf unfolding (personal communication by Simone Sabbatini, University of Tuscia).
- Regime 3 – Growing season with nonlimited water resources, from DoY 113 to DoY 166.
- Regime 4 – Drought period, from DoY 167 to DoY 252.
- Regime 5 – Growing season with nonlimited water resources, from DoY 253 to DoY 306.
- Regime 6 – Winter/dormant season, from DoY 307 to DoY 366.

The sensitivity of the MI model to the regime detection procedure is low. In particular, for both ADL and PADL models, the presence of lagged endogenous variables and the cubic time trend entering the model during daytime period help to drive the imputation model, even in the case that breakdates are not correctly identified. The largest negative impact on the quality of the imputations is mainly attributable to the presence of long gaps crossing two regimes, and such impact could even be more accentuated in the case of abrupt changes (e.g. cutting).

[4] Thomas, C.K, Law, B.E., Irvine, J., Martin, J.G., Pettijohn, J.C., Davis, K.J. (2009). Seasonal Hydrology Explains Interannual and Seasonal Variation in Carbon and Water Exchange in a Semiarid Mature Ponderosa Pine Forest in Central Oregon. Journal of Geophysical Research 114, G04006. doi:10.1029/2009JG001010

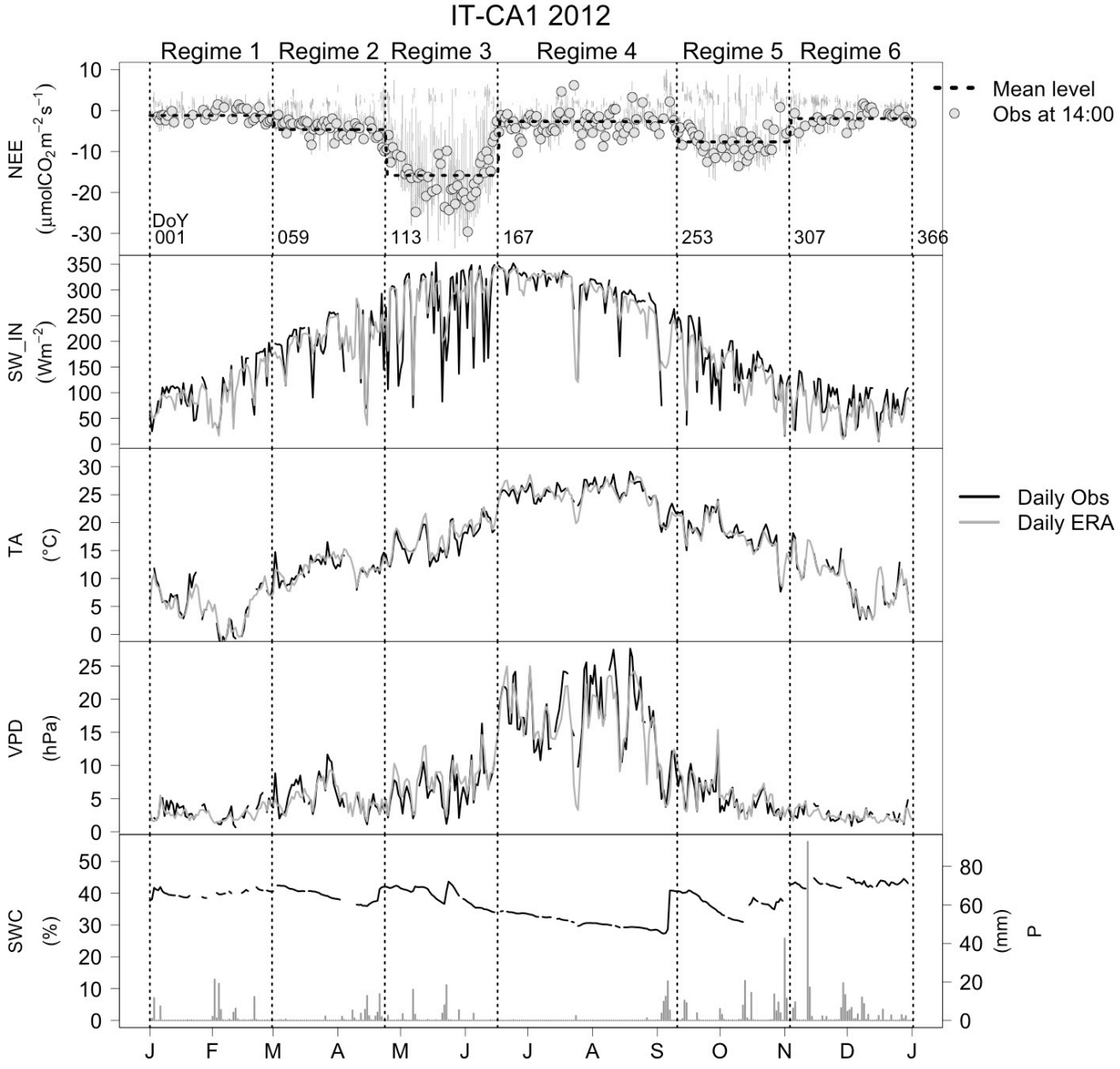

Figure A2. An example of regime detection for the IT-CA1 use case.

Appendix B: Daytime and nighttime separation and MAR assumption.

As we said in Subsection 2.5, the data were further split by daytime and nighttime periods according to a global radiation threshold of 10 Wm^{-2}. In the case of the PADL model, the daytime and nighttime subdivision was performed by evaluating the median diurnal cycle of the global radiation in each regime, in order to ensure that time series length T was constant for each cross-sectional data point. Figure B1 shows the daytime and nighttime periods for the IT-CA1 use case.

Performing the analysis separately for daytime and nighttime periods is important not only to distinguish between different ecophysiological processes (that is, CO_2 assimilation processes during daytime, and CO_2 release from the ecosystem to the atmosphere through respiration processes during nighttime), but has also the essential function of making the MAR assumption more likely. In fact, even if there is no way to test whether the MAR hypothesis holds in a given data set, the number of NEE missing data points is higher during nocturnal regimes, when low-turbulence conditions cause many data to be rejected by the quality control routines (e.g. u^*-filtering procedure). In addition, because of the absence of photosynthetic activity, NEE data during nighttime are likely to assume positive values, as they only reflect the ecosystem respiration processes. This means that if data are not analysed separately, the probability of being missing depends on the missing values themselves, and it is higher for nocturnal data points. If we could know the true value of the missing data, and if we knew that such value were positive, there would be a high chance of predicting exactly that the missing data point had occurred during the nighttime. In this case, it cannot be excluded that the missingness pattern is non ignorable, and therefore that proper MI procedures cannot be devised. When data are, instead, analyzed separately by daytime and nighttime subperiods, there is no reasonable way to relate the probability of missingness to the values assumed by NEE. This behavior is not specifically valid, of course, for NEE only and, *mutatis mutandis*, the same considerations apply equally well to other flux variables.

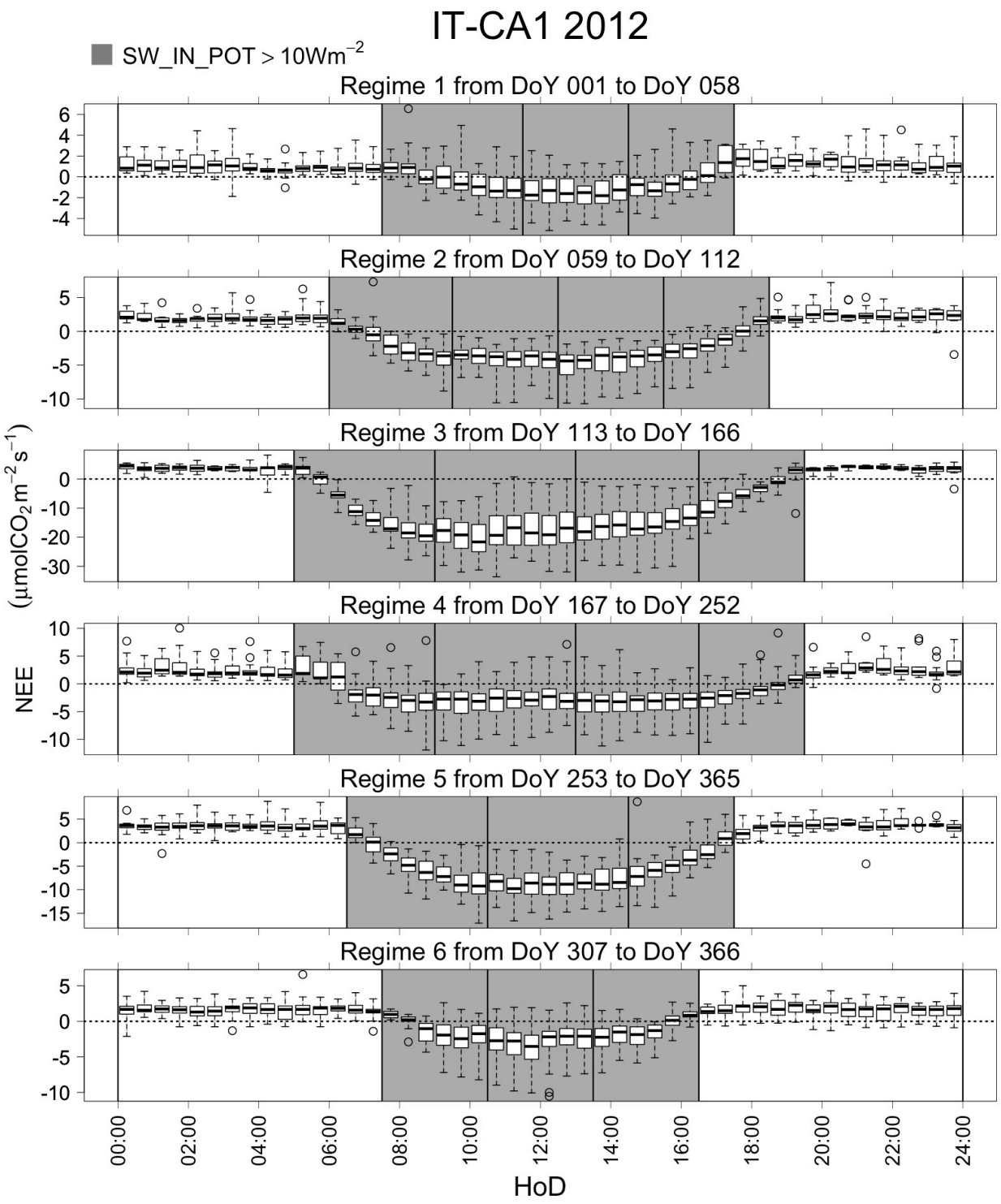

Figure B1. An example of subdivision into daytime and nighttime subsets for the PADL model for the IT-CA1 use case.

Table 1S. Annual budget estimates, associated uncertainties, and fraction of missing information ρ of Latent Heat (LE, MJ $m^{-2}y^{-1}$) gap-filled flux, reconstructed through MDS and PADL algorithms.

Site ID / Year	Model	$\tilde{Q}$	Uncertainty (MJ $m^{-2}y^{-1}$)					ρ
			Within	Between	Total	95% CI		
		MJ $m^{-2}y^{-1}$	$\bar{U}^{1/2}$	$B^{1/2}$	$2\tilde{V}^{1/2}$	Lower	Upper	
AT-Neu 2010	MDS	1018	16		32	987	1050	
	PADL	1007	16	3	33	974	1040	0.04
AU-Cpr 2012	MDS	756	9		18	738	774	
	PADL	787	9	2	18	768	805	0.04
AU-How 2011	MDS	2403	27		53	2350	2456	
	PADL	2475	27	10	57	2418	2533	0.12
DK-Sor 2009	MDS	1182	21		42	1141	1224	
	PADL	1189	21	2	42	1146	1231	0.01
FI-Hyy 2007	MDS	852	11		23	829	875	
	PADL	854	12	2	23	831	877	0.02
FR-Pue 2008	MDS	1068	13		26	1042	1094	
	PADL	1069	13	2	27	1042	1095	0.02
GF-Guy 2008	MDS	3399	35		71	3328	3470	
	PADL	3483	36	13	76	3407	3559	0.11
IT-CA1 2012	MDS	1430	17		33	1397	1464	
	PADL	1424	17	3	35	1389	1459	0.03
US-Los 2006	MDS	589	10		20	569	609	
	PADL	585	10	1	20	565	606	0.00
US-Ne2 2012	MDS	1776	35		70	1706	1846	
	PADL	1799	35	3	70	1729	1869	0.01

Table 2S. Annual budget estimates, associated uncertainties, and fraction of missing information ρ of Sensible Heat (H, MJ $m^{-2}y^{-1}$) gap-filled flux, reconstructed through MDS and PADL algorithms.

Site ID / Year	Model	$\tilde{Q}$	Uncertainty (MJ $m^{-2}y^{-1}$)					ρ
			Within	Between	Total	95% CI		
		(MJ $m^{-2}y^{-1}$)	$\bar{U}^{1/2}$	$B^{1/2}$	$2\tilde{V}^{1/2}$	Lower	Upper	
AT-Neu 2010	MDS	96	9		18	78	113	
	PADL	88	9	4	20	67	108	0.15
AU-Cpr 2012	MDS	2169	32		65	2104	2233	
	PADL	2187	32	2	65	2123	2252	0.00
AU-How 2011	MDS	1294	22		44	1250	1337	
	PADL	1284	22	4	45	1238	1329	0.04
DK-Sor 2009	MDS	61	22		44	17	105	
	PADL	40	22	1	44	-4	85	0.00
FI-Hyy 2007	MDS	495	18		36	459	530	
	PADL	519	18	3	36	483	555	0.02
FR-Pue 2008	MDS	617	24		48	568	665	
	PADL	708	24	3	49	660	757	0.02
GF-Guy 2008	MDS	763	17		35	729	798	
	PADL	751	18	4	36	714	787	0.05
IT-CA1 2012	MDS	1285	27		55	1230	1340	
	PADL	1332	27	2	55	1277	1387	0.01
US-Los 2006	MDS	810	16		32	778	843	
	PADL	803	16	2	33	770	836	0.01
US-Ne2 2012	MDS	737	27		54	684	791	
	PADL	727	27	3	54	673	781	0.01

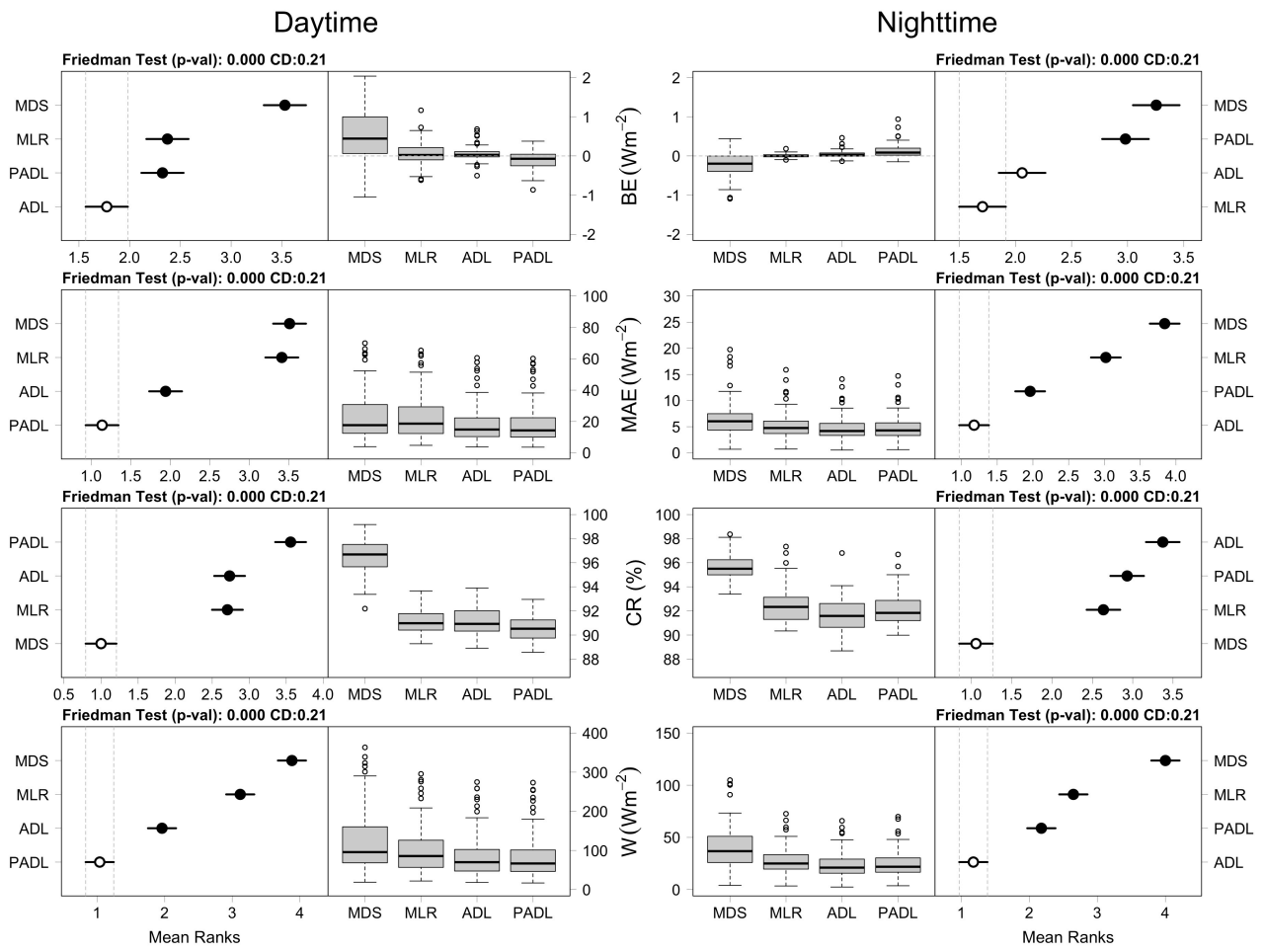

Figure 1S. Average in-sample bias error (BE, Wm^{-2}), mean absolute error (MAE, Wm^{-2}), coverage rate (CR, %) and 95% confidence interval width (W, Wm^{-2}) of Latent Heat (LE) gap-filled flux, for the MDS algorithm and the three proposed MI models, separately for daytime and nighttime subsets (defined according to a global radiation threshold of 10 Wm^{-2}). The right and left panels show a graphical visualization of Nemenyi test, where dots represent the average ranks for each imputation model, while horizontal lines represent the confidence interval (CI) built on the basis of the critical difference (CD) value. Dashed vertical grey lines indicate the CI associated with the best ranking method. For any imputation model such that its average rank is outside these bounds and the Friedman ANOVA is significant (filled dot), there is evidence of significant difference in the mean performance with respect to the best method.

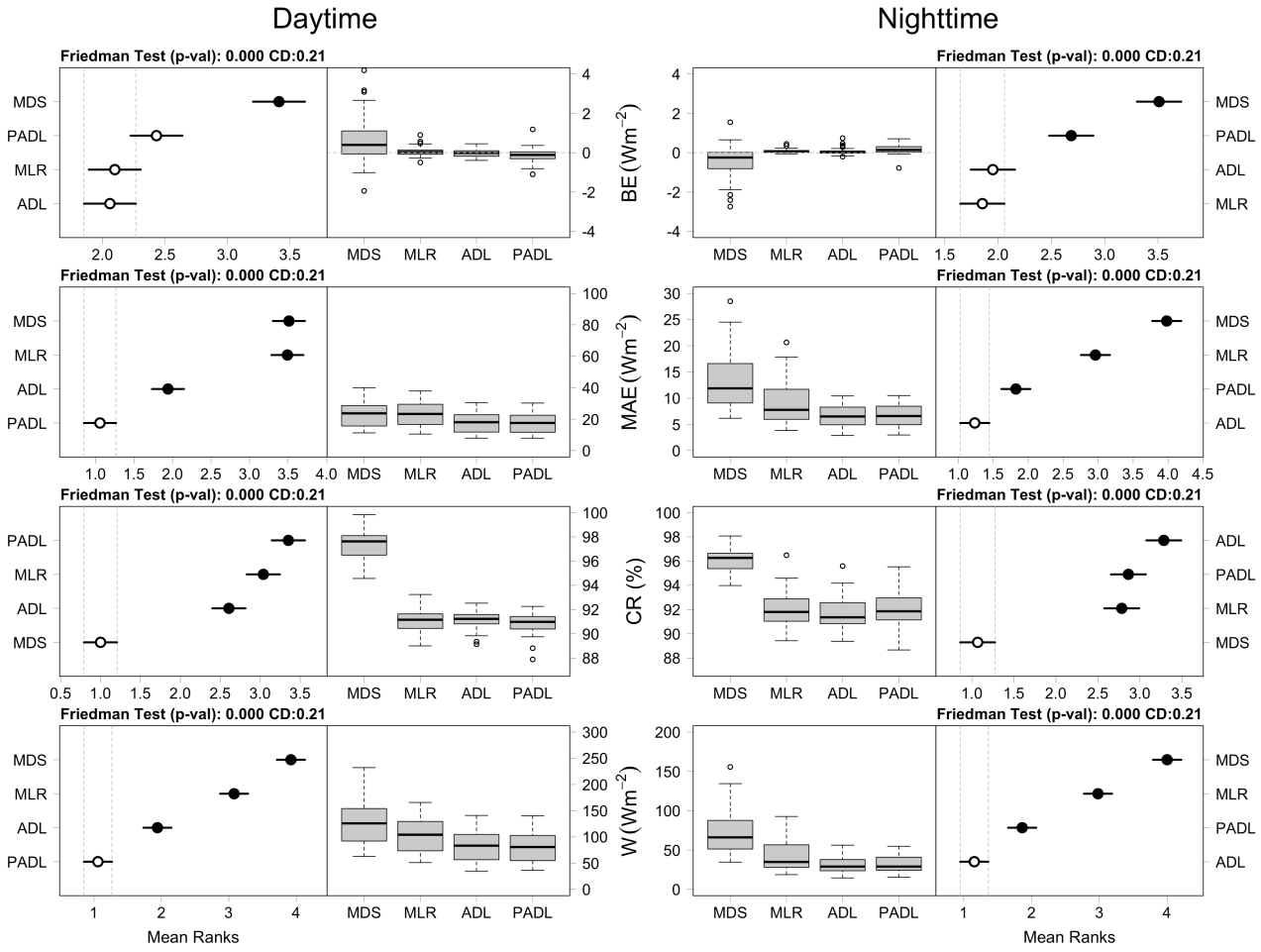

Figure 2S. Average in-sample bias error (BE, Wm^{-2}), mean absolute error (MAE, Wm^{-2}), coverage rate (CR, %) and 95% confidence interval width (W, Wm^{-2}) of Sensible Heat (H) gap-filled flux, for the MDS algorithm and the three proposed MI models, separately for daytime and nighttime subsets (defined according to a global radiation threshold of 10 Wm^{-2}). The right and left panels show a graphical visualization of Nemenyi test, where dots represent the average ranks for each imputation model, while horizontal lines represent the confidence interval (CI) built on the basis of the critical difference (CD) value. Dashed vertical grey lines indicate the CI associated with the best ranking method. For any imputation model such that its average rank is outside these bounds and the Friedman ANOVA is significant (filled dot), there is evidence of significant difference in the mean performance with respect to the best method.

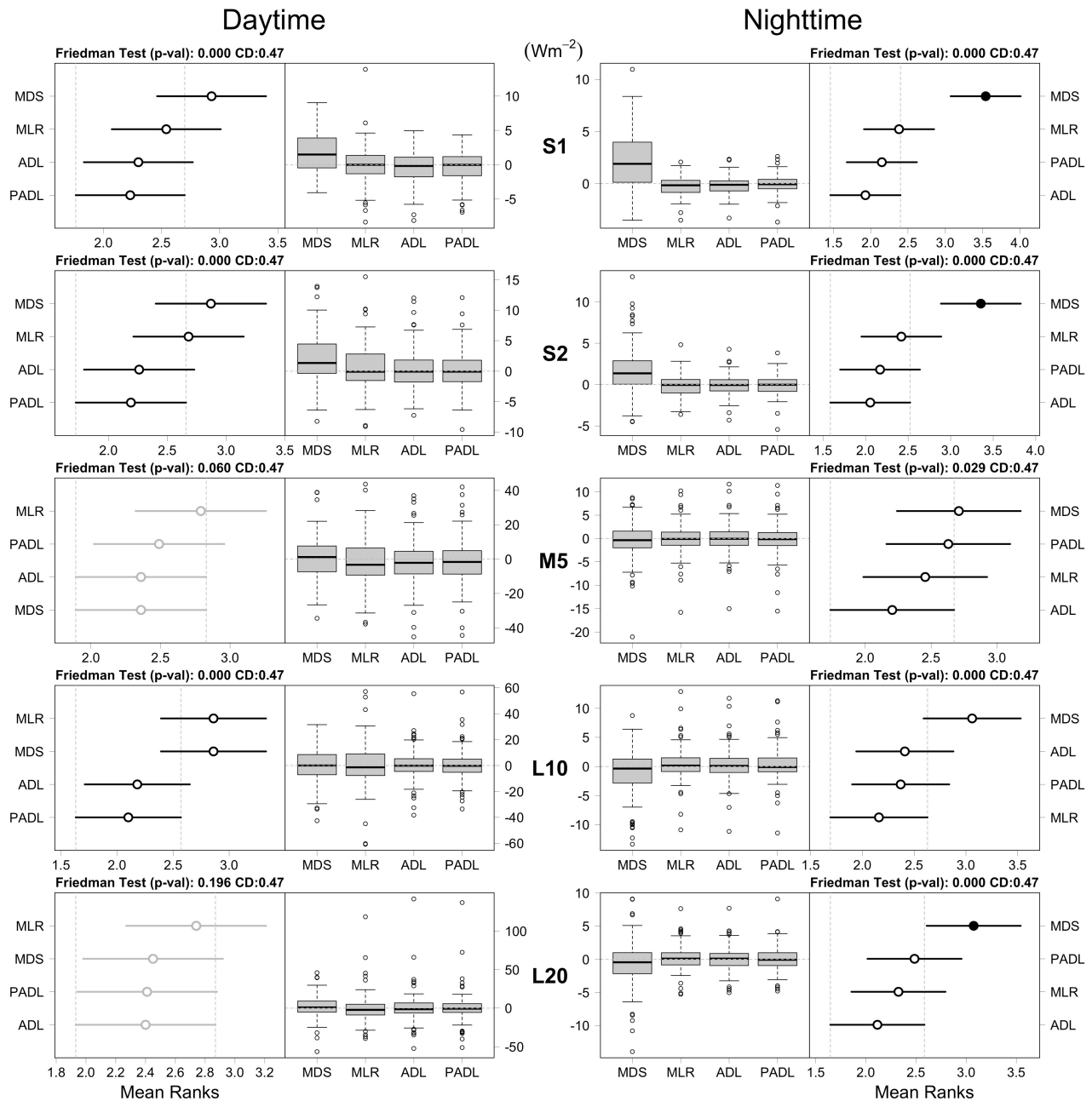

Figure 3S. Comparison of the out-of-sample bias error (BE) of Latent Heat (LE, Wm^{-2}) gap-filled flux, for the MDS algorithm and MLR, ADL and PADL imputation models, in the 5 synthetic gap scenarios and separately for daytime and nighttime subsets. The boxplots under each scenario were obtained by 10 simulations for each benchmark site, giving a total of 100 simulations for each imputation model. The right and left panels show a graphical visualization of Nemenyi test, where dots represent the average ranks for each imputation model, while horizontal lines represent the confidence interval (CI) built on the basis of the critical difference (CD) value. Dashed vertical grey lines indicate the CI associated with the best ranking method. For any imputation model such that its average rank is outside these bounds and the Friedman ANOVA is significant (filled dot), there is evidence of significant difference in the mean performance with respect to the best method.

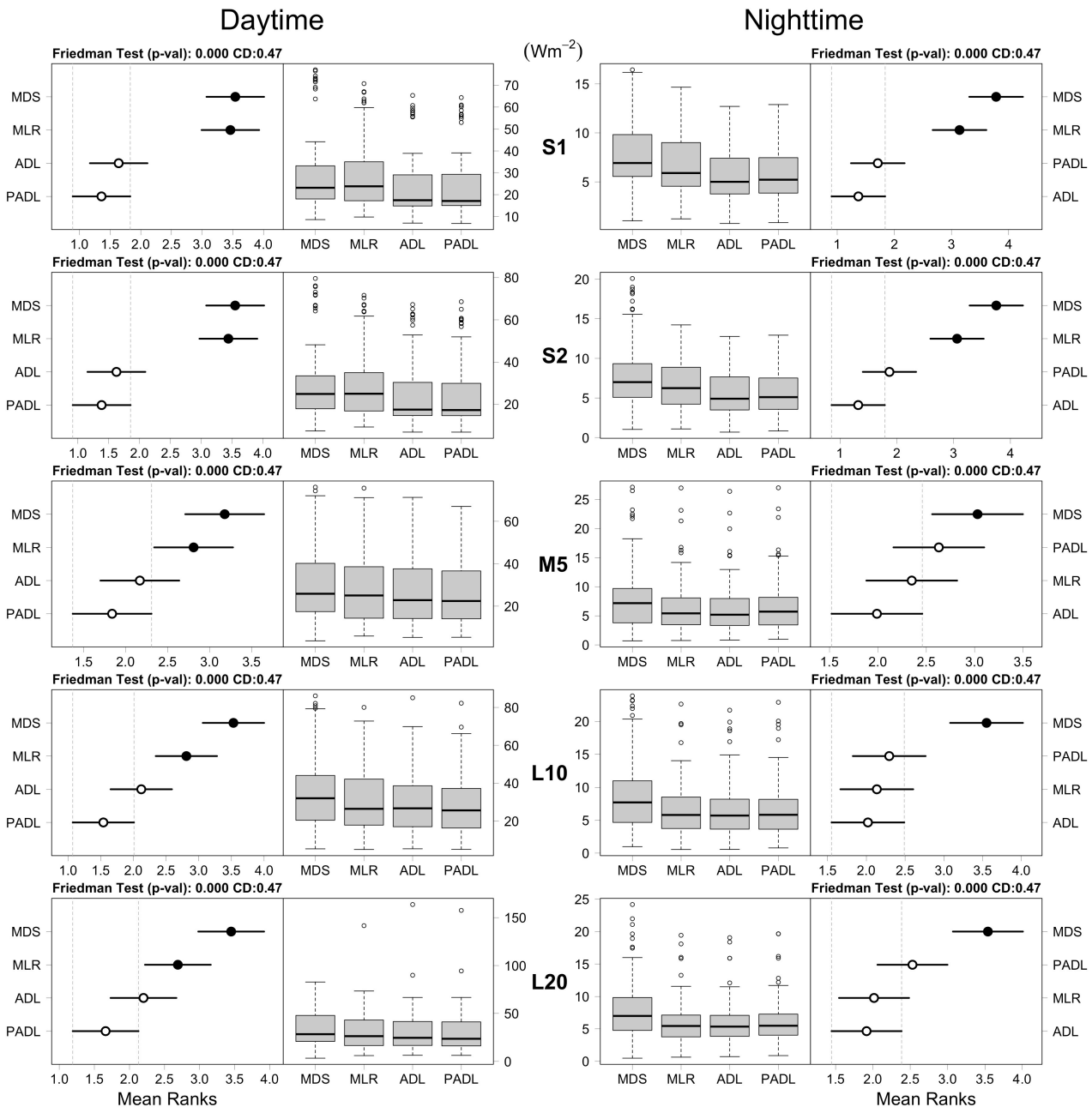

Figure 4S. Comparison of the out-of-sample mean absolute error (MAE) of Latent Heat (LE, Wm^{-2}) gap-filled flux, for the MDS algorithm and MLR, ADL and PADL imputation models, in the 5 synthetic gap scenarios and separately for daytime and nighttime subsets. The boxplots under each scenario were obtained by 10 simulations for each benchmark site, giving a total of 100 simulations for each imputation model. The right and left panels show a graphical visualization of Nemenyi test, where dots represent the average ranks for each imputation model, while horizontal lines represent the confidence interval (CI) built on the basis of the critical difference (CD) value. Dashed vertical grey lines indicate the CI associated with the best ranking method. For any imputation model such that its average rank is outside these bounds and the Friedman ANOVA is significant (filled dot), there is evidence of significant difference in the mean performance with respect to the best method.

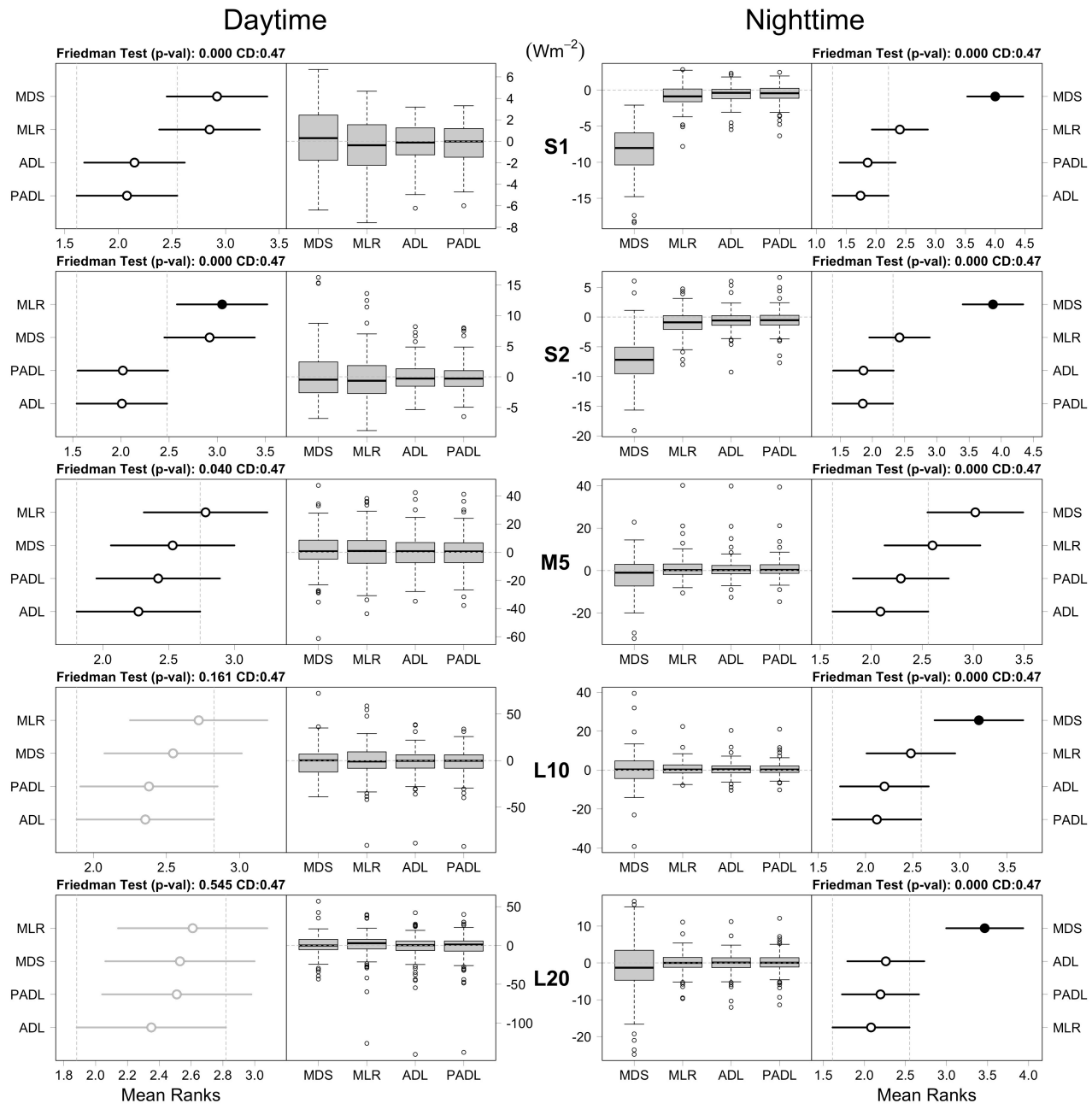

Figure 5S. Comparison of the out-of-sample bias error (BE) of Sensible Heat (H, Wm^{-2}) gap-filled flux, for the MDS algorithm and MLR, ADL and PADL imputation models, in the 5 synthetic gap scenarios and separately for daytime and nighttime subsets. The boxplots under each scenario were obtained by 10 simulations for each benchmark site, giving a total of 100 simulations for each imputation model. The right and left panels show a graphical visualization of Nemenyi test, where dots represent the average ranks for each imputation model, while horizontal lines represent the confidence interval (CI) built on the basis of the critical difference (CD) value. Dashed vertical grey lines indicate the CI associated with the best ranking method. For any imputation model such that its average rank is outside these bounds and the Friedman ANOVA is significant (filled dot), there is evidence of significant difference in the mean performance with respect to the best method.

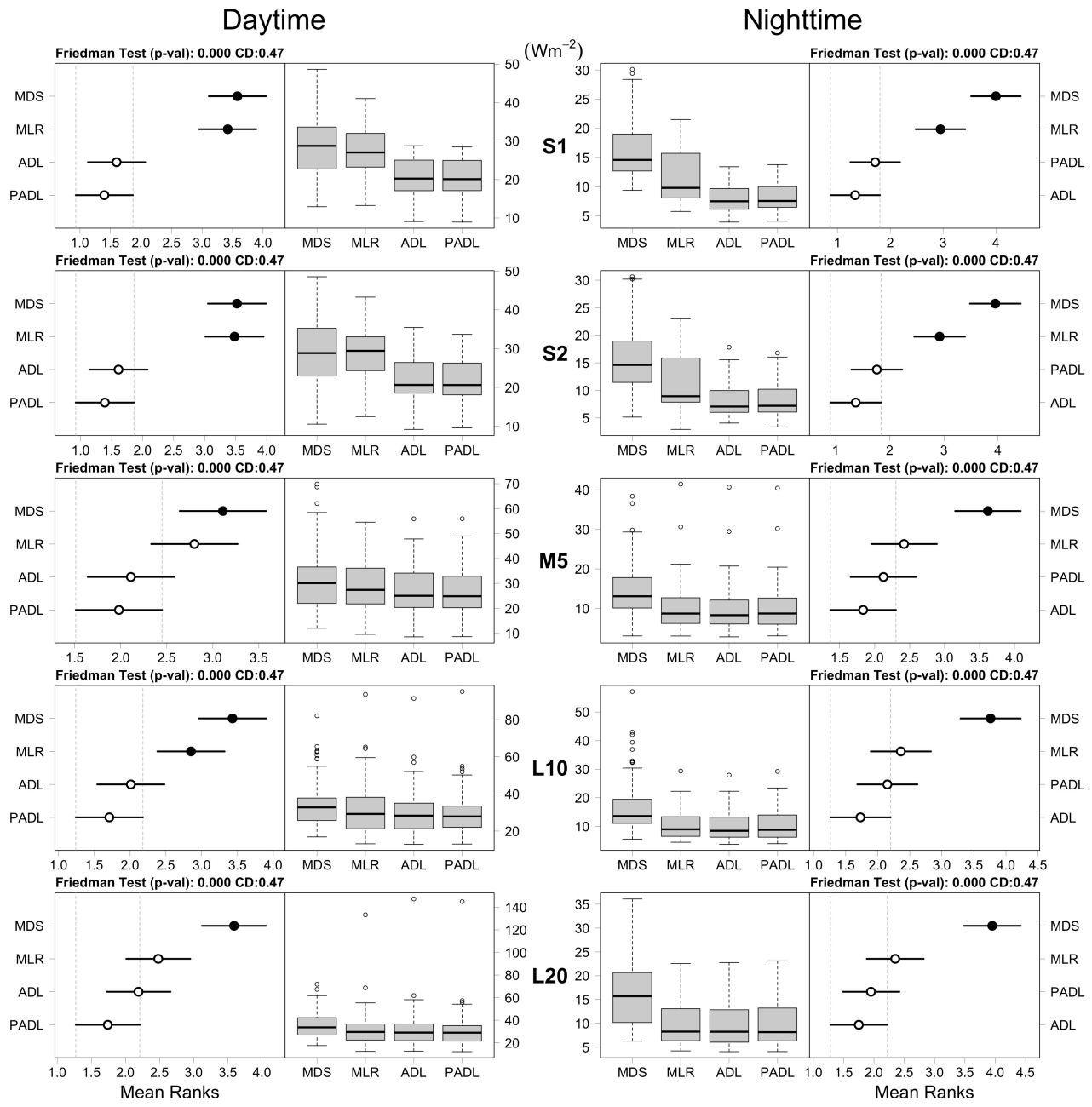

Figure 6S. Comparison of the out-of-sample mean absolute error (MAE) of Sensible Heat (H, Wm^{-2}) gap-filled flux, for the MDS algorithm and MLR, ADL and PADL imputation models, in the 5 synthetic gap scenarios and separately for daytime and nighttime subsets. The boxplots under each scenario were obtained by 10 simulations for each benchmark site, giving a total of 100 simulations for each imputation model. The right and left panels show a graphical visualization of Nemenyi test, where dots represent the average ranks for each imputation model, while horizontal lines represent the confidence interval (CI) built on the basis of the critical difference (CD) value. Dashed vertical grey lines indicate the CI associated with the best ranking method. For any imputation model such that its average rank is outside these bounds and the Friedman ANOVA is significant (filled dot), there is evidence of significant difference in the mean performance with respect to the best method.