Verification report for forecast and historical weather data

1 Verification summary

Four different meteorological parameters (air temperature, wind speed, precipitation and dewpoint temperature) have been validated at more than 10'000 different meteorological stations worldwide during the year 2017, by analysing the accuracy of several different 'stand-alone' weather forecast models, satellite observations and reanalysis models. Additionally, the accuracy of different multi-model approaches was tested and compared against 'stand-alone' models and a 24h forecast from model output statistics (MOS). Both historical and forecast data sets (models) were compared, as distinguished by the temporal availability of the model data.

The **2 m air temperature** is best modelled by the meteoblue Learning Multi-Model (MLM) with values of MAE = 1.2 K. A MOS air temperature forecast performs as well as the reanalysis model ERA5 (MAE = 1.5 K), which is recommended for historical data sets. The 'stand-alone' global weather forecast models perform in the range between 1.7 and 2.2 K. Hence, the 6-day forecast of the meteoblue Learning Multi-Model is as good as the 1-day forecast of a 'stand-alone' numerical weather forecast model.

The model uncertainty of the forecasted 10 m **wind speed** is within 1.5 – 1.7 m s^{-1} by using 'stand-alone' weather forecast models and 1.5 m s^{-1} for historical data when using the reanalysis model ERA5. The model error could be reduced to 1.2 m s^{-1} for model simulations with MOS.

The model skill of **daily precipitation events** decreases with increasing precipitation intensity. Numerical weather forecast models are recommended for small precipitation events. For heavy precipitation events, the model skill of satellite observations is higher than those of numerical weather forecast models. The model skill for daily precipitation events could not be increased by a multi-model approach mixing two (or more) models (MM). For historical data, annual **precipitation sums** are reproduced best by using satellite observations from CHIRPS2, which are bias corrected with the same data set used in this study. As a result, the accuracy of CHIRPS2 in regions without measurements is unknown and can be expected to be significantly worse, because no correction can be performed in those areas.

The accuracy of the **dewpoint temperature** is slightly worse than the accuracy of the air temperature. MAE values are between 1.9 – 2.4 K for numerical weather forecast models and 1.6 K for the reanalysis model. The accuracy of model simulations with MOS are in a similar range to those of the reanalysis model.

Table 1: Comparison of the mean absolute error (MAE) for four different meteorological parameters.

	Model approach	Air temperature	Wind speed	Annual precipitation	Dewpoint temperature
Forecast	meteoblue Learning Multi-Model (MLM)	1.2 K	–	170 mm	–
	MOS	1.5 K	1.2 m s^{-1}	–	1.7 K
	Weather forecast models	1.7 – 2.2 K	1.5 – 1.7 m s^{-1}	220 – 230 mm	1.9 – 2.4 K
History	Reanalysis model	1.5 K	1.5 m s^{-1}	120 – 180 mm	1.6 K

Table 2: Recommendations for historical analysis and operational forecast configuration:

	History	Forecast
Air temperature	ERA5	meteoblue Learning Multi-Model (MLM)
Wind speed	ERA5	meteoblue MOS and meteoblue model mix
Precipitation (events)	ERA5 (all precipitation events) CMORPH (heavy precipitation events)	meteoblue Learning Multi-Model (MLM)
Precipitation (annual sums)	meteoblue historical model mix	meteoblue Learning Multi-Model (MLM)
Dewpoint temperature	ERA5	meteoblue MOS and meteoblue model mix

2 Spatial Analysis of air temperature

Truly remarkable is the fact that the meteoblue Learning Multi-Model (MLM), which computes actual forecast outperforms the best historical reanalysis (ERA5). As a comparison to a reanalysis model (ERA5) and to the best forecast (MLM) the results of 'stand-alone' weather forecast models (e.g. GFS, NEMS) are shown (Figure 3). The 'stand-alone' model output of meteoblue NEMS is significantly better than GFS, but it is clear that the raw model data performs significantly worse than MOS, the reanalysis model ERA5 and MLM in particular.

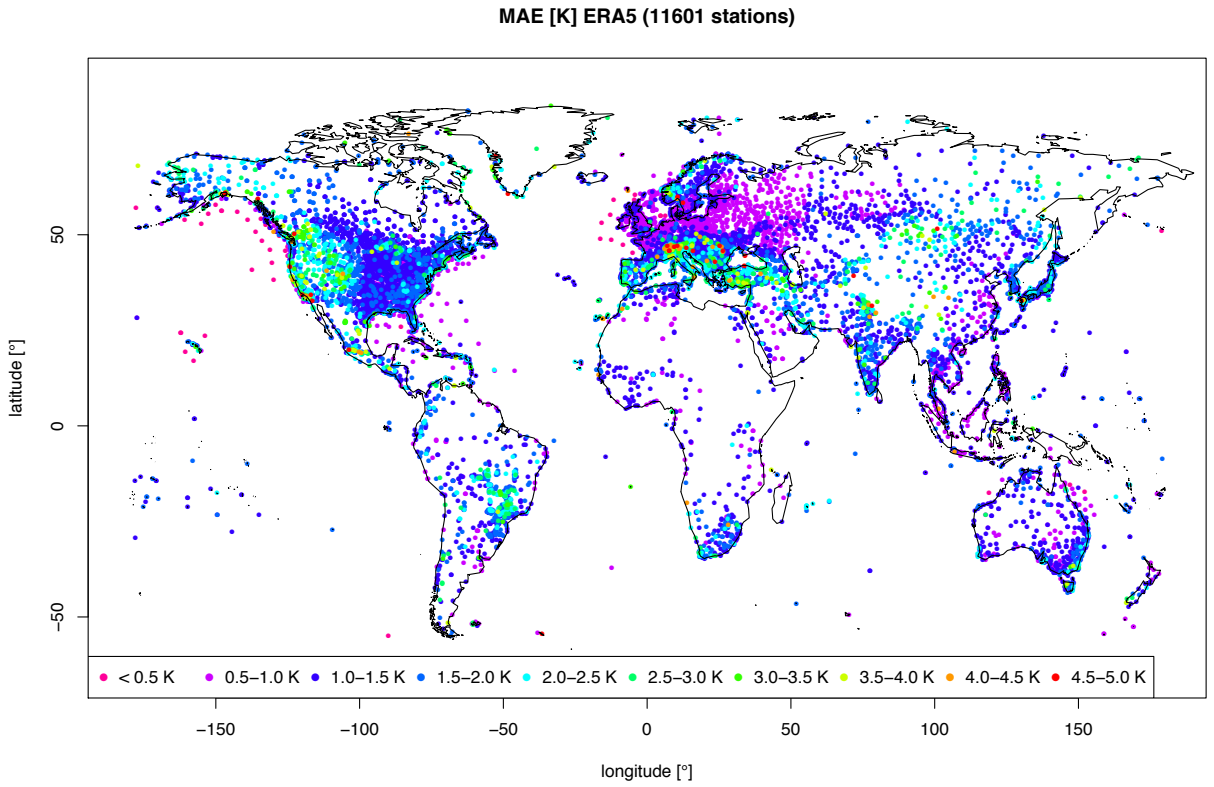

Fig. 1: MAE [K] of the 2 m air temperature of the ERA5 reanalysis (not available as forecast) used for long term **historical** analysis. Verification is based on all hourly data of the year 2017.

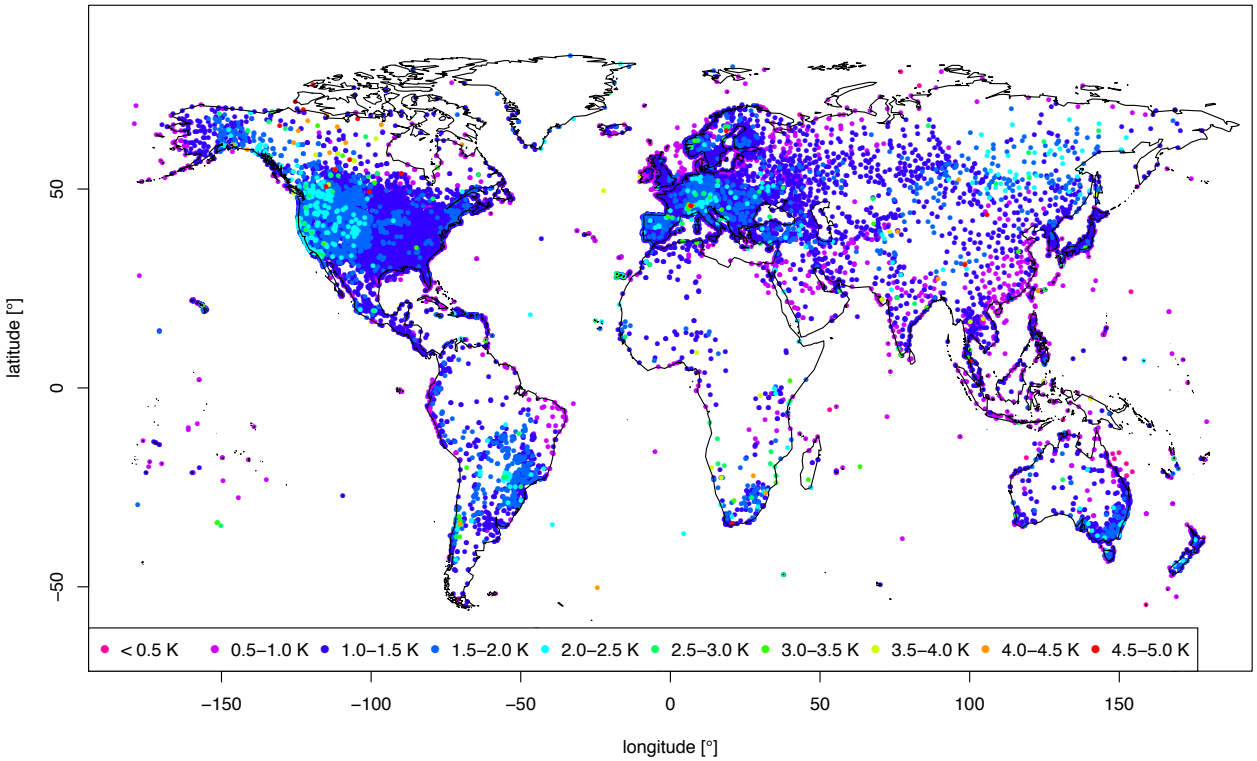

Fig. 2: MAE [K] of the 2 m air temperature of the meteoblue Learning Multi-Model (MLM) used in operational weather **forecast**. Verification is based on all hourly data from September and October 2018.

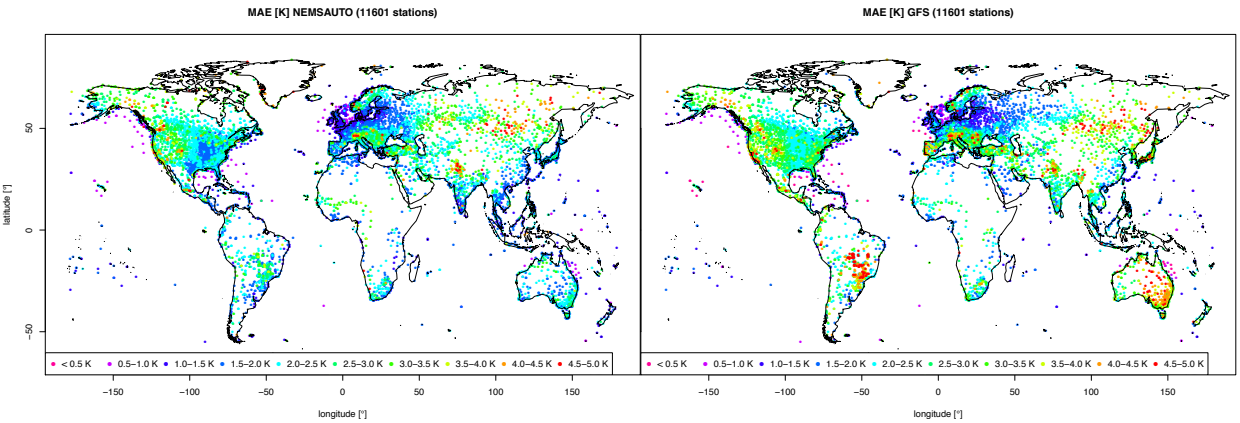

Fig. 3: MAE [K] of the 2 m air temperature of the 'stand-alone' model output as computed by meteoblue NEMS (left) and GFS (right), respectively. Verification is based on all hourly data of 2017.

3 Spatial analysis of wind speed

The ERA5 **historical** reanalysis for the 10 m wind speed is of very similar quality as the **operational** meteoblue multi-model mix. Numerical weather forecast models have significantly larger values of the MAE than ERA5, MOS or the meteoblue multi-model mix.

Table 3: Mean absolute error (MAE), mean bias error (MBE), root mean square error (RMSE) and standard deviation (stddev) [m s^{-1}] and Pearson correlation coefficient for numerical weather forecast models (GFS, NEMS), the meteoblue multi-model mix and historical reanalysis model ERA5.

	MAE [m s^{-1}]	MBE [m s^{-1}]	RMSE [m s^{-1}]	stddev [m s^{-1}]	Pearson correlation
ERA5	1.49	0.03	1.93	1.62	0.66
meteoblue multi-model mix	1.48	0.13	1.94	1.66	0.67
GFS	1.69	0.24	2.20	1.87	0.58
NEMS	1.67	-0.05	2.20	1.85	0.60

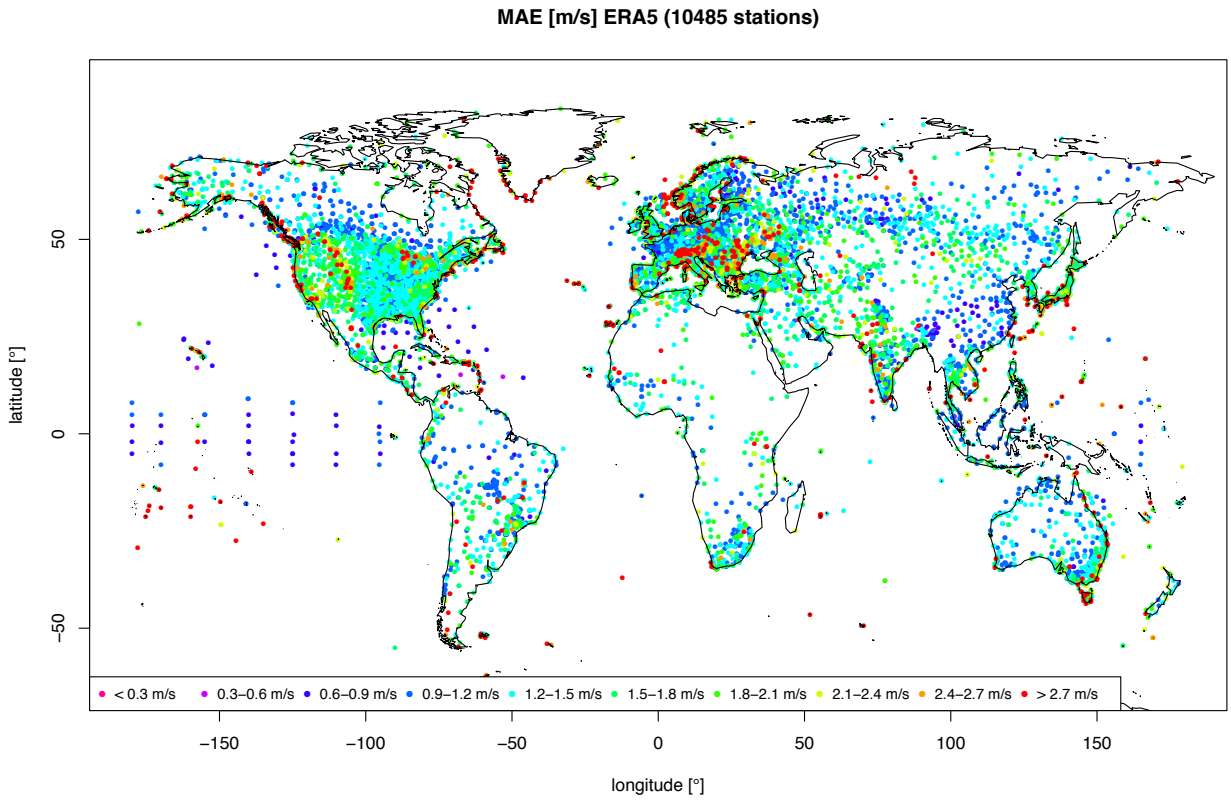

Fig. 4: MAE [m s^{-1}] of the 10 m wind speed of the reanalysis model ERA5 (not available as forecast) used for long term **historical** analysis. Verification is based on all hourly data of the year 2017.

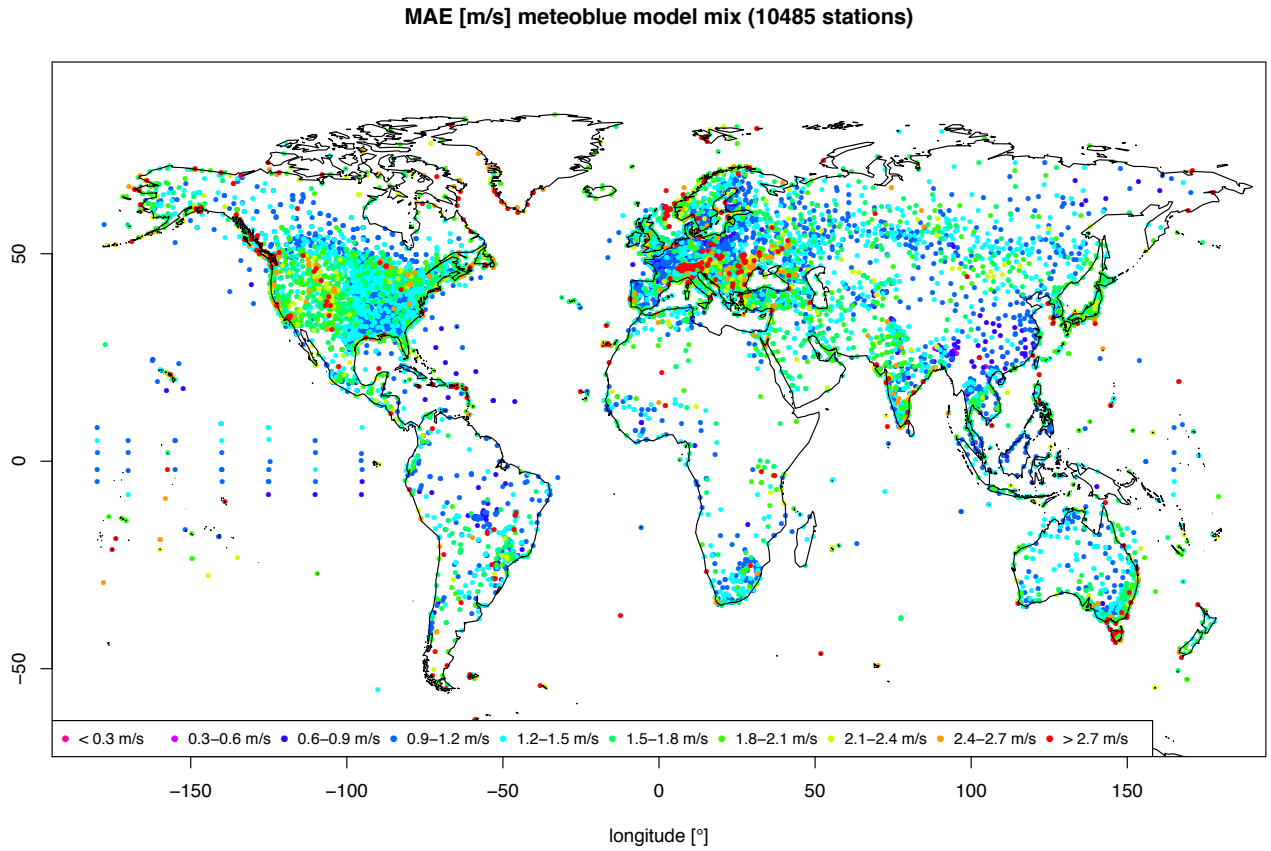

Fig. 5: MAE [m s^{-1}] of the 10 m wind speed of the meteoblue multi-model mix used in operational weather **forecast**. Verification is based on all hourly data of the year 2017.

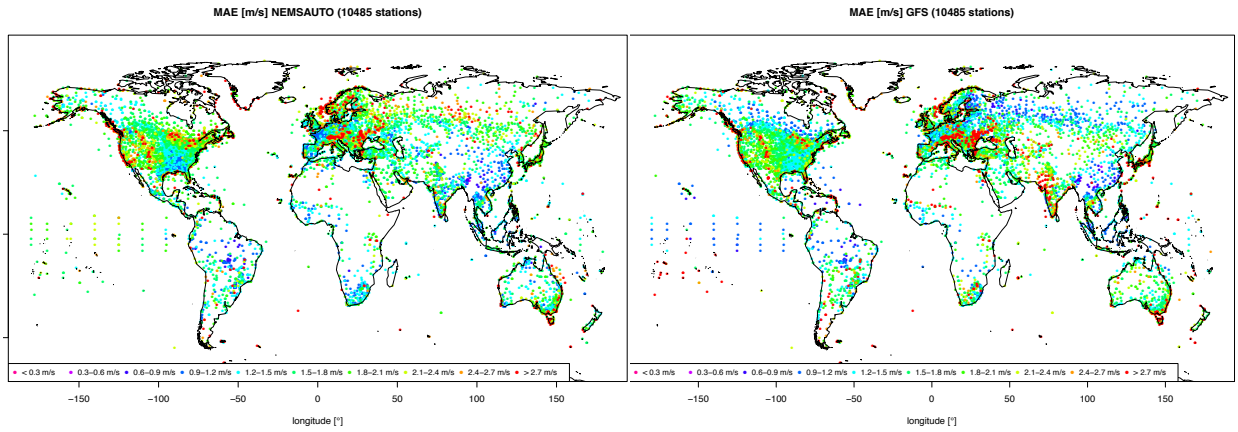

Fig. 6: MAE [m s^{-1}] of the 10 m wind speed of the 'stand-alone' model output as computed by meteoblue NEMS (left) and GFS (right), respectively. Verification is based on all hourly data of the year 2017.

4 Spatial Analysis of precipitation

For historical data, the accuracy of ERA5 and the meteoblue Learning Multi-Model (MLM) is significantly better than the satellite observation CHIRPS2 (Table 4). Satellite observations have a larger skill than numerical weather forecast models for heavy precipitation and close to the equator. For annual precipitation sums the meteoblue multi-model mix is significantly better than ERA5. Note, that the measured precipitation sums in Romania are inaccurate, resulting in non-reliable values of the mean percentage error (MPE) in this region (Figure 9).

Table 4: Probability of detection (POD), false alarm rates (FAR) and Heidke skill score (HSS) for three different daily precipitation events (1 mm; 10 mm; 50 mm) for the historical reanalysis model ERA5, the numerical weather forecast model GFS, the satellite observation CHIRPS2 and the meteoblue multi-model.

	Daily precipitation > 1 mm			Daily precipitation > 10 mm			Daily precipitation > 50 mm		
	POD	FAR	HSS	POD	FAR	HSS	POD	FAR	HSS
ERA5	0.69	0.51	0.45	0.43	0.64	0.35	0.11	0.76	0.14
GFS	0.69	0.54	0.42	0.40	0.69	0.30	0.09	0.83	0.12
CHIRPS2	0.41	0.55	0.30	0.42	0.69	0.31	0.18	0.79	0.19
meteoblue Learning multi-model (MLM)	0.70	0.49	0.47	0.48	0.64	0.36	0.09	0.73	0.14

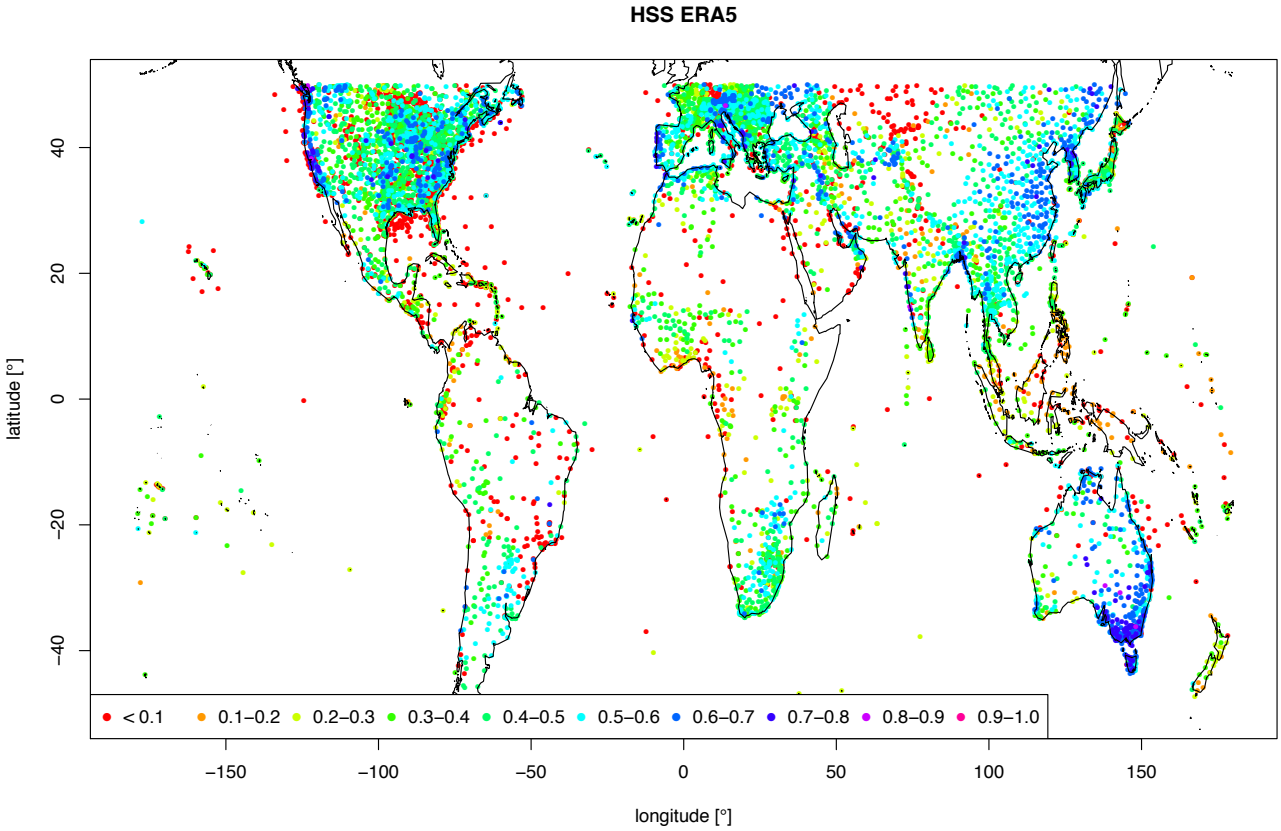

Fig. 7: Heidke Skill Score (HSS) for precipitation events of >1mm/day for the reanalysis model ERA5 (not available as forecast) used for long term **historical** analysis. Verification is based on all daily data of the year 2017.

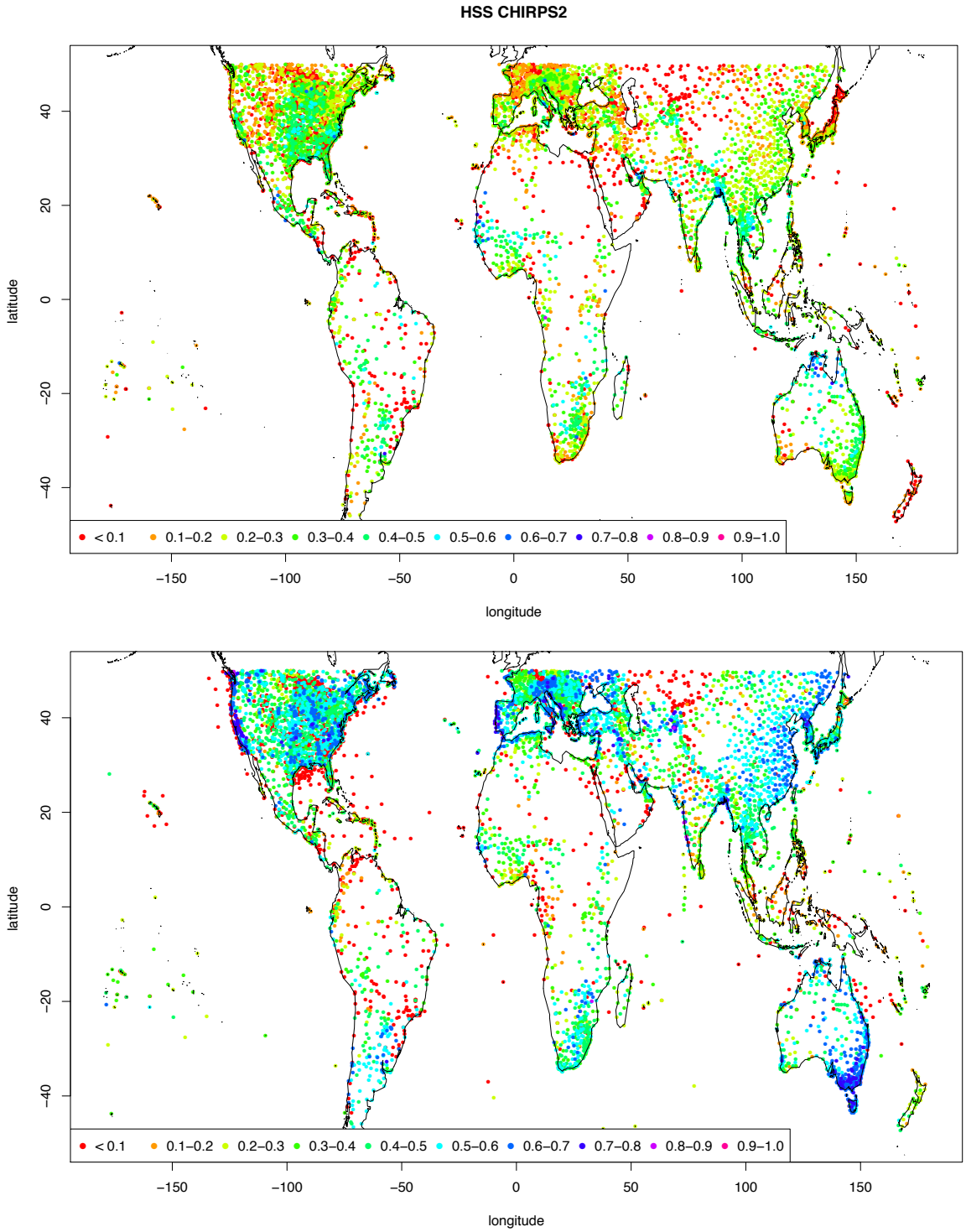

Fig. 8: Heidke Skill Score (HSS) for precipitation events of >1mm/day for the satellite observation CHRIPS2 (not available as forecast) used for long term **historical** analysis (top) and the meteoblue Learning Multi-Model (MLM) used in **forecasts** (bottom). Verification is based on all daily data of the year 2017.

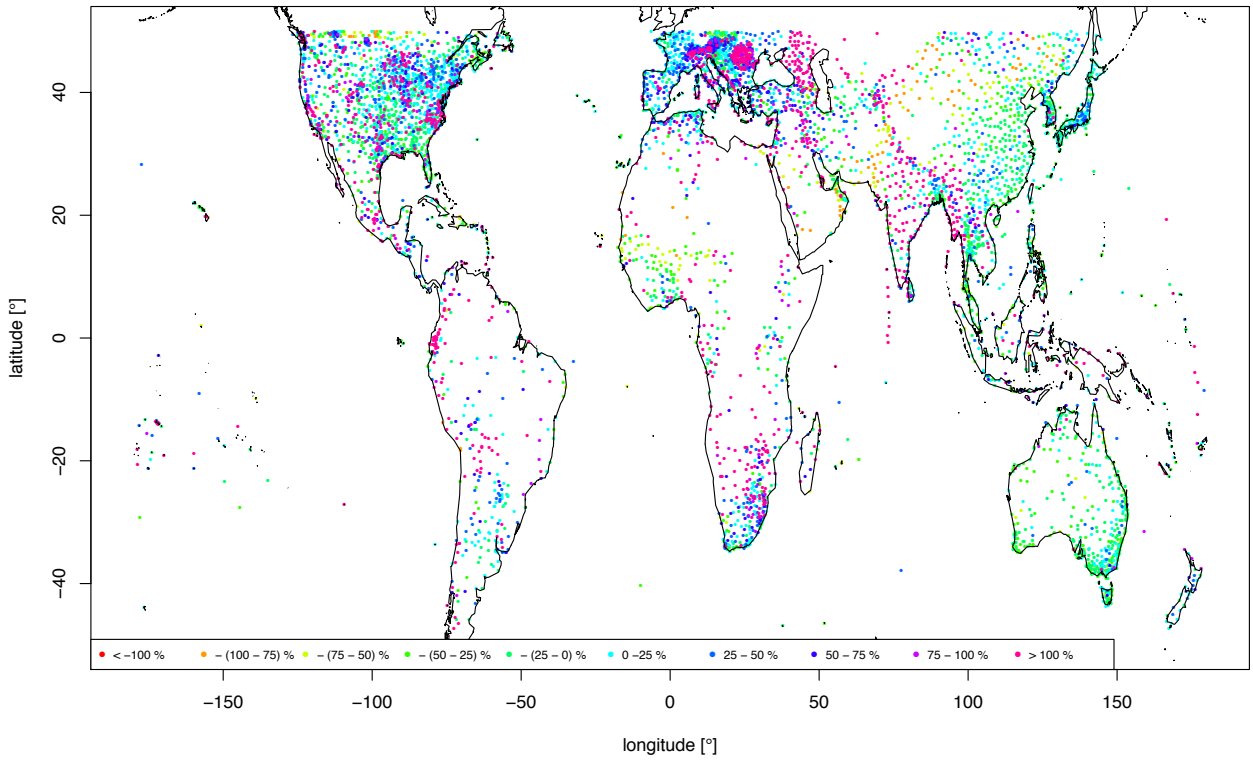

Fig. 9: MPE [%] for the meteoblue Learning Multi-Model (MLM) used in operational forecasting. Verification is based on all daily data of the year 2017.

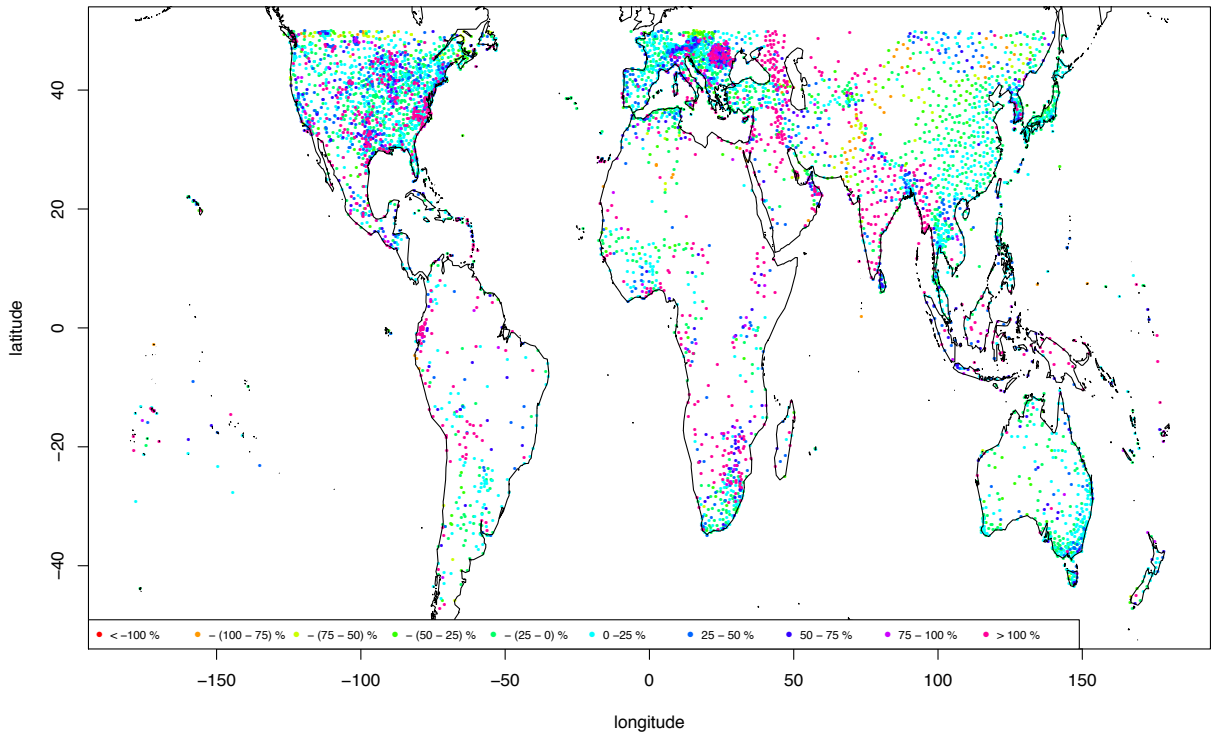

Fig. 10: MPE [%] for the **historical** meteoblue multi-model mix. Verification is based on all daily data of the year 2017.

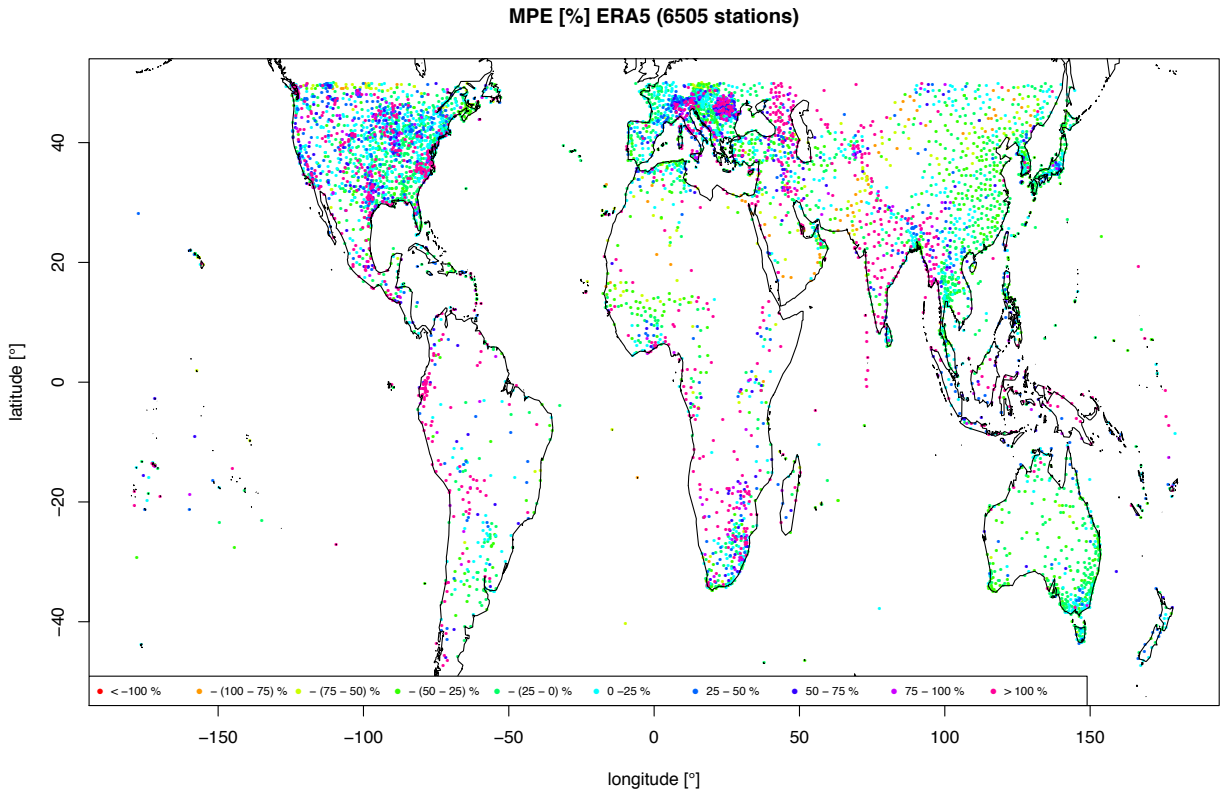

Fig. 11: MPE [%] for the **historical** reanalysis model ERA5. Verification is based on all daily data of the year 2017.

5 Spatial Analysis of the dewpoint temperature

The accuracy of the meteoblue **forecast** multi-model mix (MAE = 1.8 K) is in a similar range as the ERA5 **historical** reanalysis model (MAE = 1.6 K). The spatial distribution of the values of the MAE can be found in Figure 12 and Figure 13. The accuracy of the reference numerical weather forecast model GFS can be found in Figure 14.

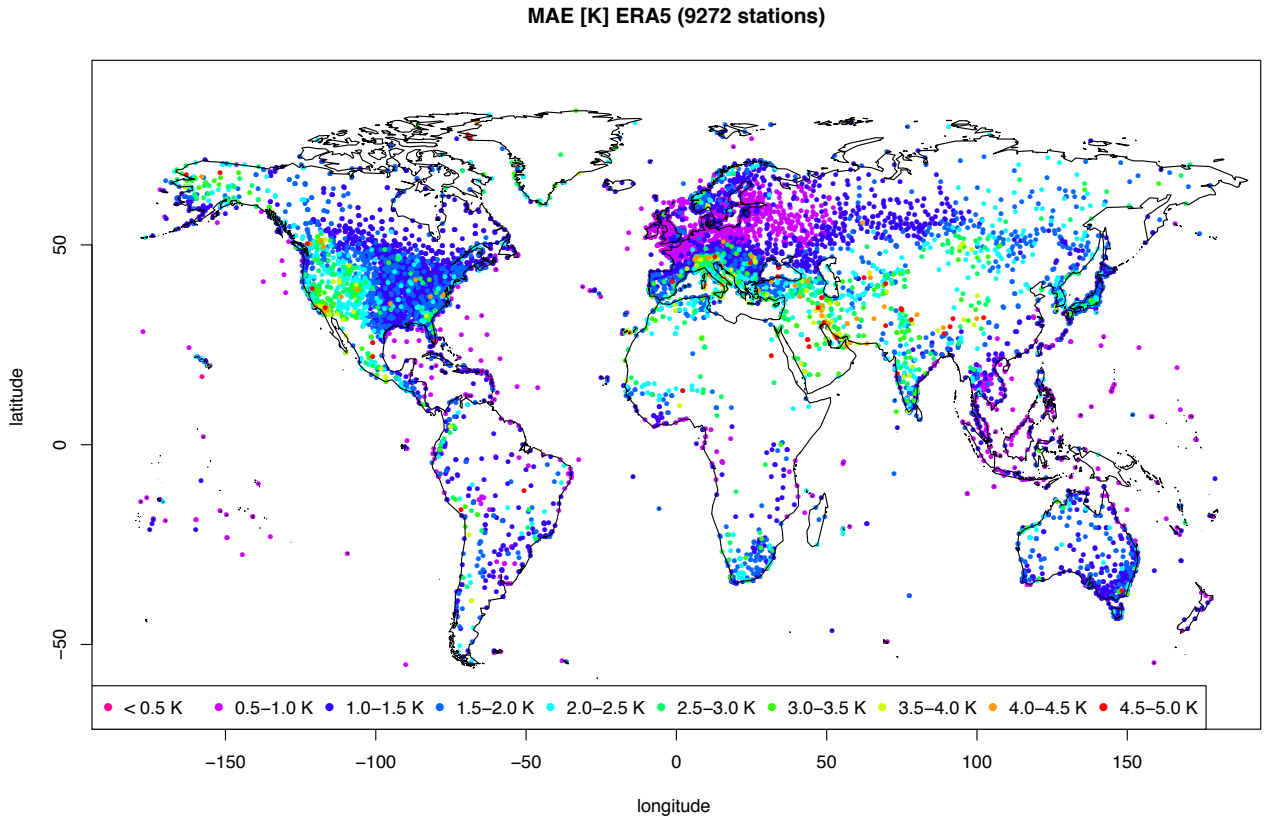

Fig. 12: MAE [K] of the reanalysis model ERA5. Verification is based on all hourly data of the year 2017.

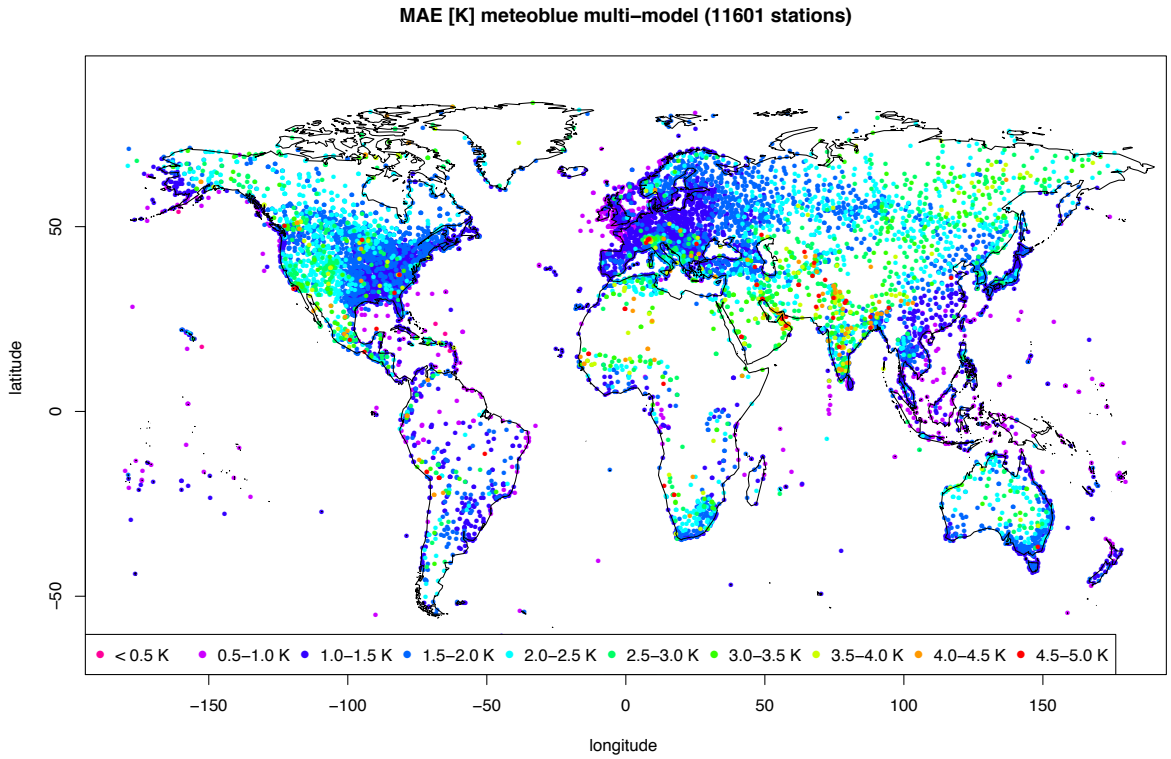

Fig. 13: MAE [K] of the meteoblue multi-model mix used in operational forecast. Verification is based on all hourly data of the year 2018.

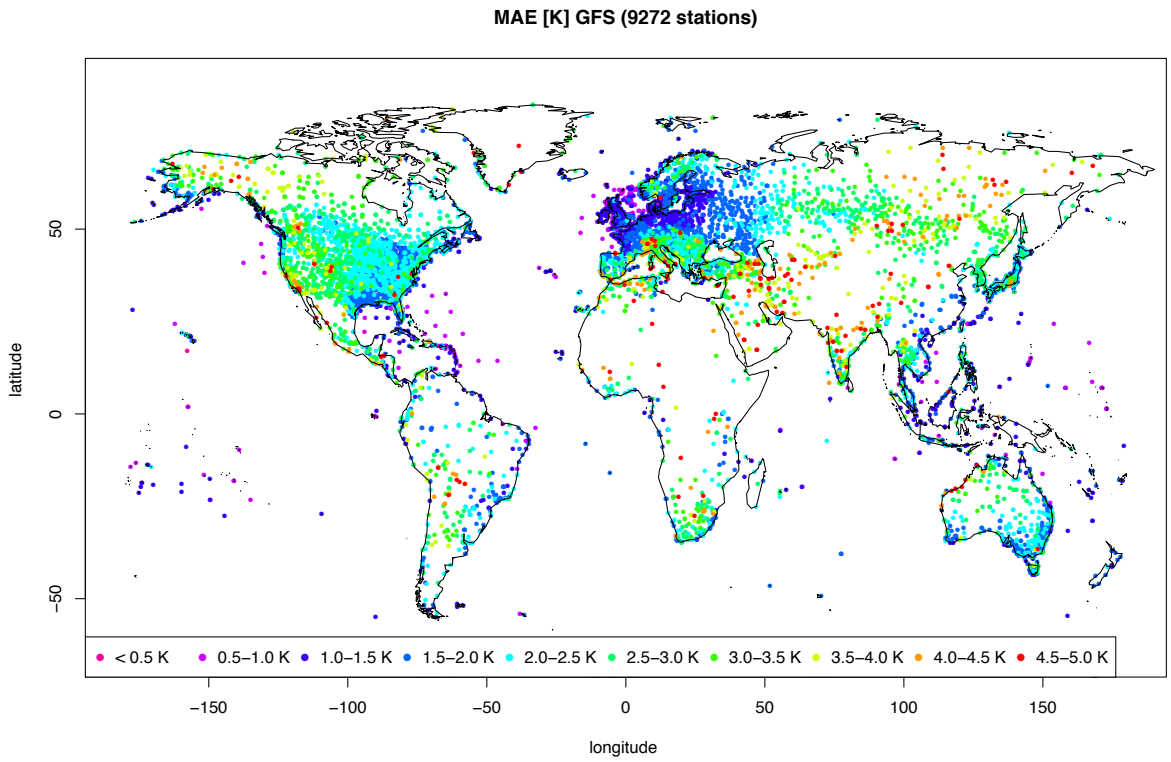

Fig. 14: MAE [K] of the numerical weather forecast model GFS. Verification is based on all hourly data of the year 2017.