IAC Institute for Atmospheric and Climate Science

Master Thesis

Comparison of COSMO-TERRA and COSMO-CLM^2 in weather mode for summer heat extremes

Author:

Verena Bessenbacher

Master Programme Atmospheric and Climate Sciences (D-ERDW)

bverena@student.ethz.ch, 16-930-604

Supervisors:

Prof. Dr. Sonia Seneviratne, Head of Land-Climate Dynamics, IAC, ETH, *sonia.seneviratne@ethz.ch*

Dr. Edouard Léopold Davin, Land-Climate Dynamics, IAC, ETH *edouard.davin@env.ethz.ch*

Dr. Jean-Marie Bettems, APND, MeteoSwiss, *jean-marie.bettems@meteoswiss.ch*

Date: May 31, 2018

ETH Zürich

Institute for Atmospheric and Climate Sciences

Universitätsstrasse 16

8092 Zürich

Acknowledgements

Picture on titlepage by Johannes Senn, 48,3625313, 8,2463943

Thanks to my supervisors for their help, encouragement and dedication
Thanks to Yiftach Ziv and Matthieu Leclair for the simulations
Thanks to Joel, John, Christoph, Sarah, Judith, Moritz for their insightful comments
Thanks to D26.1 for the endless supply of schoggi

Abstract

The benefits of coupling the atmospheric model COSMO to an advanced land surface model has been demonstrated in the context of climate simulations. Especially for mid-latitude heat waves happening on the time-scale of several weeks, which are strongly amplified by soil moisture temperature feedbacks, land-atmosphere interactions play a key role. Correctly representing radiative and surface fluxes during such events is thus of crucial importance for reliable weather forecasting and risk prevention.

This study aims to further investigate the added value such of an advanced surface representation in the context of numerical weather simulations of summer heat extremes. For this purpose, (1) COSMO with the land surface model TERRA and (2) COSMO coupled to the Community Land Model (CLM) are compared to assess their performance in the anomalously hot European summers of 2003, 2006 and 2015.

Overshooting sensible heat production at the expense of evapotranspiration in COSMO-TERRA is consistent among all summers and induces a 1-2 K warm bias in COSMO-TERRA. Additionally, with increasing vegetation density, COSMO-TERRA shows decreasing performance in latent heat estimation, calling for a review of the evapotranspiration parameterisations. COSMO-CLM suffers from biases of the same direction but smaller magnitude as COSMO-TERRA, but its results are more aligned with evaluation datasets. A simple statistical benchmark outperforms both models in latent heat estimation. The superior surface flux representation in COSMO-CLM, however, does not translate into temperature. COSMO-CLM temperatures are especially hampered by a consistent 2-3 K cold bias in radiative temperatures, most severe for daily maximum temperature. The extremeness of the three summers, in contrast, is represented best in COSMO-CLM. Indices measuring the danger of the extreme conditions for human health show overall large differences among indices in dangerous conditions, calling for further investigation.

COSMO-CLM is disadvantaged by sharing the parameter tuning of COSMO-TERRA. Furthermore, an outdated aerosol climatology in COSMO introduces a known cold bias especially in the Mediterranean. A further investigation of both models with an improved aerosol climatology and a tuning tailored to COSMO-CLM could alleviate performance of COSMO-CLM and could answer remaining questions. In addition, it would be interesting to assess whether the recent developments of the new COSMO-TERRA version 5.05 could partially remove identified deficiencies in the standard COSMO-TERRA version.

In conclusion, this master thesis has shown the importance of land surface processes for the correct prediction of summer hot extremes at the weather time scale. The results could help illuminate potential pathways towards the development of better predictions of such events in Europe.

Contents

List of Figures

List of Tables

1 Introduction

Summer heat extremes pose a considerable threat to European citizens. A prolonged heatwave not only causes immediate effects such as heat-related deaths and health risks (Kovats and Hajat, 2008), but also intermediate effects on infrastructure and agriculture (e.g. Fink et al., 2006, COPA COGECA, 2003, respectively). The record summer heat in 2003 (Beniston, 2004, Schär and Jendritzky, 2004), for example, caused a total excess mortality of more than 70,000 deaths (Robine et al., 2008) and a economic loss of 13 Billion Euros across Europe (De Bono et al., 2004).

With the projected impact of regional climate change, an exacerbation of heat extreme events is looming on the horizon: European summer heat spells will likely increase in frequency, intensity and duration by the end of the 21st century (Beniston et al., 2007, Koffi and Koffi, 2008). The intensity of extreme temperatures increases more than a simple shift of mean temperatures would suggest (Beniston et al., 2007). Seneviratne et al. (2016) found a 40% stronger warming for extreme temperatures in the Mediterranean compared to global average temperature change. Especially in the densely populated Mediterranean region and in low-level southern Europe, frequency and duration of heat extremes increase faster and health-related impacts are more severe than in northern parts of Europe (Fischer and Schär, 2010).

The disproportionally high change in extreme temperatures over Europe and the increase in heat extreme frequency can potentially be explained by an increased variability in interannual temperatures (Schär et al., 2004, Vidale et al., 2007), which is mainly attributed to land-atmosphere feedbacks (Seneviratne et al., 2006, Vidale et al., 2007).

1.1 The Role of Land Processes in Regional Climate

The land surface interacts through several processes with the atmosphere, affecting regional weather and climate. The albedo of the ground determines the amount of shortwave radiation absorbed by the ground. Absorbed energy is stored and gradually released back to the atmosphere through warming of air (sensible heat), emission of longwave radiation and evapotranspiration of water (latent heat). The energy budget at the surface can thus be described with

$$SW_{net} + LW_{net} = LH_{net} + SH_{net} + GHF_{net} \tag{1}$$

with net shortwave radiation SW_{net} and net longwave radiation LW_{net} determining the amount of energy absorbed in the ground, which is divided into net latent heat flux LH_{net}, net sensible

heat flux SH_{net} and ground heat flux GHF_{net} (in the following without the subscript) by processes taking place in the soil and vegetation.

Heat storage in the ground is important on diurnal time scales, where part of the absorbed energy is stored during day and released during night, dampening diurnal temperature ranges. This effect similarly takes place on seasonal timescales.

The latent heat flux links the energy and water budget of the land surface. Through stomatal conductance and assimilation rates, vegetation and soil moisture content determine the amount of latent heat released into the atmosphere. The soil acts as a sink for precipitation and source for evapotranspiration. The slow-changing nature of soil moisture leads to the soil moisture memory effect, where integrated precipitation over several months determines the current soil moisture content. In a transitional climate regime, where evapotranspiration is not limited by available energy, soil moisture constrains evapotranspiration (Seneviratne et al., 2010), hence the partition between sensible and latent heat and thus has an influence on local temperature.

Through this soil-moisture temperature feedback, the land surface is coupled with the atmosphere. Low soil moisture distorts the ratio of incoming radiation fed back to the atmosphere towards less latent heat and more release of sensible heat in transitional soil moisture regimes (Ferranti and Viterbo, 2006). Therefore, a precipitation deficit in spring can amplify summer temperatures by degrading the soil moisture and allowing for more sensible heat to be released by the surface in summer. This coupling is especially high during large-scale anticyclonic regimes, where atmospheric circulation is low (Quesada et al., 2012), and when the soil moisture regime is transitional (Seneviratne et al., 2010). Quesada et al. (2012) followed that amplification of heat extremes through the soil-moisture temperature feedback is conditional of two features: (1) when anticyclonic weather prevails and (2) dry conditions in winter and spring degrade soil moisture content, shifting local conditions to a transitional soil moisture regime.

Simulations have shown an amplification of heat extremes by strong soil moisture coupling in most of the recent European heatwaves (Fischer et al., 2007a, Jaeger and Seneviratne, 2011). With climate change, transitional soil moisture regimes are expected to cover also northern areas of Europe (Seneviratne et al., 2006) and projected soil moisture drying increases extreme temperatures over Europe in simulations until 2100 (Seneviratne et al., 2013).

The importance of land surface processes for heat extremes over Europe demonstrates the urgency for european weather services to correctly include land-atmosphere feedbacks into their numerical weather prediction models to assure precise forecast and representation of heat extremes.

1.2 Land Surface Models and Their Current Deficiencies

However, the current numerical weather prediction model (NWP model) COSMO-TERRA of the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss) lags behind more advanced land surface schemes. For regional climate models (RCMs), the benefits of coupling an atmospheric model to an advanced land surface model (LSM) has already been demonstrated in the context of climate simulations: Davin et al. (2011) compare MeteoSwiss' COSMO-TERRA with COSMO-CLM, where the land surface model TERRA is replaced by the Community Land Model version 3.5 (CLM) for climate simulations, i.e. on lower temporal (monthly) and spatial (50 km) resolution than NWP models (Beware of the different terminology used in this study: COSMO-CLM is referred to as COSMO-CLM^2, COSMO-TERRA is COSMO-CLM). They found improved flux partition in COSMO-CLM compared to COSMO-TERRA, which particularly in summer enhanced representation of temperature, precipitation and cloud cover (Davin et al., 2011).

CLM and TERRA differ in their structural complexity of representation of land surface processes. The manifold biological and ecological processes taking place on the land surface in vegetation and soil cannot be represented in LSMs from first physical principles, but rather need careful parameterisation reasoned by sound knowledge of the processes involved. Most prominently, CLM directly simulates stomatal conductance and photosynthesis and can therefore loosely be categorised as third generation LSM (after classification from Sellers et al., 1997). In contrast, TERRA belongs to the earlier – second generation – LSMs using empirical relationships to estimate evapotranspiration. Furthermore, CLM allows for more detailed representation of the land surface, for example by explicitly modelling different vegetation types and allowing them to coexist on one grid cell.

Davin et al. (2016) show that COSMO-CLM outperforms COSMO-TERRA and other RCMs in 2-meter mean, minimum and, most pronounced, maximum temperature. They demonstrate that these improvements largely stem from a better representation of fluxes, including evapotranspiration. They furthermore conclude that improvements of land surface schemes in RCMs are likely to mend current known RCM deficiencies.

A similar study for both models on higher temporal (hourly) and spatial (6.6 km) resolution (i.e., in "weather mode") has not yet been conducted. Especially during hot summer extremes, surface fluxes are known to have a large impact on temperature (Donat et al., 2017) and better representation of surface fluxes has shown to be central for model performance improvement (Davin et al., 2011). Correctly representing the land surface in NWP models is thus both of crucial importance for reliable weather forecasting and, at the same time, a promising pathway towards substantial model performance improvements.

1.3 The Summer Heat Extremes of 2003, 2006 and 2015

2003, 2006 and 2015 list among the top ten of the record-breaking heat events over Europe since 1950 and are the subset thereof that took place in Central Europe (Russo et al., 2015). The three summers show the largest temperature anomalies across Europe since 1910, together with the 2010 Russian-centered heatwave (NOAA, 2018). In the following, they are shortly characterised with the help of Figure 1.

2003 was the hottest ever recorded summer in Central Europe (Luterbacher et al., 2004). García-Herrera et al. (2010) summarize that "an anomalously persistent northerly displacement of the Atlantic Subtropical High" led to prounounced heat waves in June and August. During the first two weeks of August, this high slowly advanced towards north-east (Figure 1), leading to record-breaking temperatures over most of Central Europe. On 14^{th} of August, for the first time since beginning of August, no temperature record was broken (García-Herrera et al., 2010). Fischer et al. (2007b) state that the warm and dry spring anomaly preceeding the event degraded soil moisture across Europe far below multiyear averages. In sensitivity experiments they find that surface temperature anomalies would have been reduced by 40% when prescribing climatological soil moisture.

The **2006** heatwave was located more northerly than the 2003 heatwave, affecting especially Germany and its neighbouring countries (Rebetez et al., 2009). It was primarily active mid-July (Figure 1), where a persistent Omega Blocking disabled zonal flow (Rebetez et al., 2009). 2006 was more anomalous than 2003 in terms of area affected, and low soil moisture is likely to have contributed to the severity of the event (Rebetez et al., 2009).

In **2015**, four heat waves from June throughout September (Sippel et al., 2016) marked the year as second-warmest in the Central European domain investigated by (Orth et al., 2016). A northward displacement of the jet stream towards the end of the summer led to a successive pattern of recorded heat waves: In south-western Central Europe, temperatures were highest beginning of June, while south-eastern and north-eastern Central Europe experienced their main heat extremes in July and August, respectively (Figure 1, Sippel et al., 2016). Throughout summer 2015 a record precipitation deficit of 31% below mean precipitation resulted in the summer being the driest since 1904. The subsequent anomalously large soil moisture deficit was lower than in any other 21st century european summer (Orth et al., 2016).

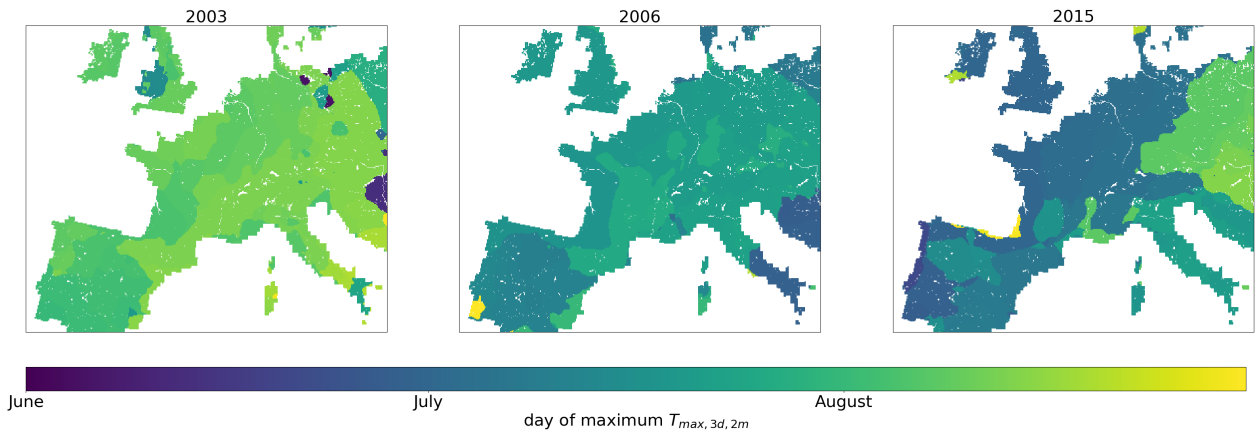


Figure 1: Day of seasonal temperature record for 2003, 2006 and 2015 expressed as day where maximum $T_{max,3d,2m}$ was reached throughout JJA of respective year in the domain of the simulations. $T_{max,3d,2m}$ is the 3-day running mean daily maximum temperature at 2-meters in the EOBS dataset.

1.4 Objectives

In the light of the severity of these heat extremes and their link to soil moisture conditions, there are ongoing efforts in the COSMO consortium to evaluate and improve land surface processes representation in COSMO in the framework of the COSMO Priority Task TERRA NOVA (see Bettems, 2017, Ziv, 2017).

However, as compared to a climate model in the study of Davin et al. (2016), in weather models the urge to improve land surface representations is perceived smaller. Since weather models are running at most a couple of days, and with the initial conditions being regularly refreshed, the state of the land surface can in principle be corrected on the basis of observations. Furthermore, in the time frame of several days, especially soil moisture with its integrated nature does not change much.

Nevertheless, we argue this argumentation oversees the fact that data assimilation is especially challenging for soil conditions, since horizontal heterogenity is high and the number of in-situ measurements low. An improved surface model is hence also beneficial to data assimilation quality and at the same time can potentially increase forecast quality of NWP models in the same order of magnitude than for RCMs shown by Davin et al. (2016).

Consequently, in this study we aim to investigate the added value of advanced surface representation to a numerical weather prediction system. For this purpose, two instances of the COSMO model are compared to assess their performance: (1) MeteoSwiss' COSMO is using the second generation LSM TERRA, and (2) COSMO-CLM with the third generation land surface scheme CLM. Both models are running for the anomalously hot European summers of 2003, 2006 and

2015 and on the high temporal (hourly) and spatial (6.6 km) resolution characteristic to NWP models.

Evaluation of weather forecast with COSMO-TERRA is usually confined to point-wise comparison with available meteorological station data. This study, in constrast, exploits increasingly available gridded datasets over the Central European domain.

The focus is on evaluating how the two models perform in calculating both surface temperatures and fluxes as compared to available observations to prioritise possible developments of COSMO-TERRA and pinpoint towards the main differences of the two models.

2 Methods

COSMO is a non-hydrostatic, fully compressible limited-area atmospheric model suitable for climatological and meteorological time scales and a broad range of spatial scales (Doms et al., 2011). It is developed and maintained by the Consortium for Small-scale Modelling (COSMO). The atmospheric model prognostically estimates horizontal and vertical wind components, pressure perturbation, temperature, specific humidity, cloud water content, turbulent kinetic energy (TKE) and the specific water content of rain and snow (Doms et al., 2011). Horizontal coordinates are defined on a rotated geographical grid and vertical coordinates are along generalized terrain-following height. A second-order leapfrog scheme is used for time integration (Doms et al., 2011).

2.1 Model Experiment

COSMO-TERRA and COSMO-CLM simulations are performed on hourly time resolution over Central and Southern Europe for a common period of November until end of August in 2003, 2006 and 2015 at a spatial resolution of 6.6 km. Modelled temperatures and fluxes are subsequently compared to evaluation datasets where available.

COSMO-POMPA version 5.0 run in hindcast mode (i.e., no data assimilation) was used. To provide lateral boundary conditions including sea surface temperatures, simulations are nested into ECMWF reanalysis data of the respective years, with a relaxation zone of 15 points in each cardinal direction – this relaxation zone is removed for analysis. Initial conditions are obtained from COSMO Reanalysis (Ziv, 2017). COSMO-CLM has arbitrary values assigned to the soil moisture, in COSMO-TERRA soil moisture is taken from MeteoSwiss operational archives at the beginning of the simulation. Therefore, all simulations start in November of the previous year to provide a two-month spin-up for soil moisture initialisation. To ensure maximum comparability between both

models, the parameter tuning of COSMO-TERRA is also used for COSMO-CLM, i. e. COSMO-CLM is not individually tuned.

2.2 TERRA vs CLM

The two models differ in their LSM. It provides the lower boundary conditions for the COSMO atmospheric model. The LSM TERRA (see Schrodin and Heise, 2001, Grasselt et al., 2008), is used in the current NWP models of the German Weather Service (DWD) and MeteoSwiss within the COSMO Consortium (see http://www.cosmo-model.org). The Community Land Model (CLM) version 4.0 is developed by the National Center for Atmospheric Research of the USA (NCAR) and considered a state-of-the-art LSM (Davin et al., 2011).

Calculating mass and energy fluxes between atmosphere and land using stability and roughness length formulations in the atmospheric model COSMO requires the knowledge of temperature and relative humidity at the ground (Doms et al., 2011). These two variables are provided by the LSM. Additionally, CLM also computes parts of the turbulence scheme. In the following, the most relevant aspects of the two LSMs regarding this thesis are outlined. Table 1 provides an overview of the points discussed below. See Doms et al. (2011) (TERRA) and Oleson and Lawrence (2013) (CLM) for a complete description of the LSMs and Davin et al. (2011) for a more detailed comparison. See Smiatek et al. (2008) and Lawrence and Chase (2007) for details on the surface input fields of TERRA and CLM, respectively.

Surface Heterogenity: TERRA has no sub-grid scale heterogenity, apart from dividing the grid cell into bare soil and vegetated area through the plant cover fraction. CLM uses a tile approach, where every grid cell can be populated by different land types (e.g. glacier, lake, vegetated land, etc.). The vegetated land type can consist of several plant functional types (PFTs) coexisting in one grid cell. 15 PFTs plus bare soil are available, with their characteristic physiology and structure, e.g. their rooting depth.

Soil properties: TERRA has 8 soil layers and a total soil depth of 15.24 m. Eight soil types are distinguished with their characteristic hydraulic properties, heat capacity and heat conductivity (data from the Food and Agriculture Organization of the United Nations (FAO) Digital Soil Map of the World). Each grid cell is inhabited by one soil type.

CLM has 15 soil layers. The first 10 soil layers are inhabited by soil specified by color, texture and organic matter density and reach down to 3.43 m. The last 5 soil layers are defined as bedrock and only thermally, not hydrologically, active. Furthermore, CLM accounts for sub-grid horizonal and vertical heterogenity of the soil (data from the International Geosphere-Biosphere Programme (IGBP) Global Soil Data Task 2000).

Table 1: Summary of the comparison of the main differences between TERRA and CLM.

Feature	TERRA	CLM
Surface Heterogenity	No tile approach	Tile approach, PFTs, vertical soil heterogenity
Soil properties	8 soil types, one per grid cell, no vertical soil heterogenity	Gradients of soil texture, color and organic matter content, subgrid and vertical heterogenity
Vegetation structure	Foliage-like vegetation represented by LAI and rooting depth	Vegetation represented by PFTs, with own temperature and interception reservoir
Stomatal conductance and Photosynthesis	Empirically estimated from radiation, soil water content, temperature and specific humidity after Dickinson (1984)	Stomatal conductance directly modelled as in Collatz et al. (1991) and photosynthesis after Farquhar et al. (1980)
Hydrology	Only gravitational drainage	TOPMODEL-based approach, soil water interaction with ground water included
Thermal Processes	8 thermally active soil layers, vegetation influence on radiative (ground) temperature neglected	15 thermally active soil layers, distinction between vegetation temperature and ground temperature
Radiative Fluxes	Estimated on grid-scale from temperature and albedo.	Distinction between canopy / surface radiative fluxes and diffuse / direct solar radiation
Turbulent Fluxes	TKE-based scheme	Monin-Obukhov similarity

Vegetation structure: Vegetation in TERRA is represented by external input fields of leaf area index (LAI) and rooting depth (data from the Global Land Cover map for the year 2000 (GLC2000)). The LSM internally prescribes a simple seasonal cycle on the LAI. Additionally, plant cover fraction per grid cell is defined. Vegetation in TERRA is of foliage-like nature and does not have its own heat budget, water budget or temperature.

In contrast, in CLM vegetation composition and structure is directly modelled. Initial fields specify the fractional cover of each land unit, the fractional cover of each PFT per gridcell, along with stem area index (SAI) and LAI per month, canopy top height, canopy bottom height and root fraction per soil layer (data from MODIS satellite products). Mechanisms including the structure (e. g., sunlit and shaded vegetation), thermal processes (e. g., accounting for vegetation and canopy temperature) or water storage (e. g., interception reservoir) of the vegetation are included.

Stomatal conductance and Photosynthesis: Vegetation in TERRA excerts biophysical control on evapotranspiration via the empirical stomatal conductance model of Dickinson (1984). Stomatal conductance is estimated by a stomatal resistance factor r_s. It depends on radiation, soil water content, temperature and specific humidity. Plant transpiration is subsequently a function of the fraction of ground covered by vegetation, potential evaporation and r_s.

CLM models stomatal conductance and photosynthesis explicitly. It employs the stomatal conductance model of Collatz et al. (1991), which was first applied in a model by Sellers et al. (1996), and photosynthetic carbon assimilation for C3 plants is modelled as in Collatz et al. (1991) (modified after Farquhar et al., 1980). C4 plants are also parameterised after Collatz et al. (1992) and Dougherty et al. (1994). Fluxes are subsequently calculated as weighted mean of all PFTs.

Hydrology: TERRA and CLM solve the Richards equation in 8 and 10 active soil layers of variable depth, respectively. The hydrological processes in both models include interception and evapotranspiration at the surface, infiltration into the soil, capillary and gravitational transport between soil layers as well as subsurface and surface runoff. A comprehensive snow modeling is also included, but not covered here.

While in TERRA water drained from the last soil layer disappears from the model, in CLM a groundwater model is included allowing for interaction between groundwater and soil moisture. The TOPMODEL approach based on Beven et al. (1984) for runoff parameterisation used in CLM takes water table depth and topography into account.

Thermal Processes: Both TERRA and CLM solve the heat conduction equation between all soil layers. In TERRA, heat capacity depends on the soil type and the water content, but heat

conductivity only on the average water content of the soil. The lower boundary condition is a prescribed constant climatological value, and on the first soil layer the atmospheric forcing is applied. The surface temperature T_{rad} is the mean temperature of the first soil layer, solved with the heat conduction equation. Foliage temperature is assumed to be equal to surface temperature and vegetation does not influence the surface energy balance, nor does it shade the ground. T_{2m} is calculated to compare with observations.

Heat conductivity and capacity in CLM depend on soil texture, soil organic content and water content. CLM distinguishes between ground temperature T_{rad} – the result of solving the heat conduction equation for the uppermost soil layer – vegetation temperature T_v, surface temperature T_s and diagnostically estimated 2-meter temperature T_{2m}. Vegetation temperature T_v is iteratively solved from the radiation budget. T_s is defined as the mean temperature in the air and stems from heat conductance from vegetation, ground and canopy air into the atmosphere.

Radiative fluxes in TERRA are estimated on grid-scale from temperature and albedo. CLM distinguishes between canopy and ground radiative fluxes. Furthermore, diffuse and direct solar radiation is differentiated, influencing availibility of light for photosynthesis in sunlit and shaded areas of the canopy.

Turbulent fluxes in TERRA are estimated with a TKE-based scheme and are simulated in the atmospheric part of the model. CLM comes with its own scheme based on Monin-Obukhov similarity which is used additionally to the one provided by COSMO.

2.3 Evaluation Datasets

The models are compared to gridded evaluation datasets of radiative temperature T_{rad} and 2-meter temperature T_{2m} in K, radiative fluxes (longwave LW and shortwave SW radiation) and surface fluxes (latent heat LH and sensible heat SH) in Wm^{-2} (see Table 2). Precipitation is also evaluated, but since model agreement is good (see Supplementary Figure A9), it is not included in the analysis.

The EOBS dataset (see Haylock et al., 2008) provides gridded estimates of daily mean, minimum and maximum temperature as well as precipitation over Europe on a 0.1° resolution and daily basis. It is interpolated from irregularly located meteorological stations across Europe.

The Satellite Land Surface Temperature dataset (SLST) provides hourly measured land surface temperature (LST) over Southern and Central Europe at a resolution of 5.5 km. The dataset is developed at MeteoSwiss (currently unpublished, for methodology see Duguay-Tetzlaff et al., 2015). With the high resolution available in the SLST dataset, that is even higher than model resolution,

Table 2: Gridded evaluation datasets used in this work, their resolution and extent in both space and time and the variables they provide.

Abbrev.	Variables	temporal resolution	extent*	spatial resolution	extent	Reference
EOBS	$T_{2m,max}$	daily	all	0.1 °	Europe	Haylock et al. (2008)
	$T_{2m,mean}$		all			
	$T_{2m,min}$		all			
SLST	T_{rad}	hourly	all	5 km	C. Europe, Africa**	similar to Duguay-Tetzlaff et al. (2015)
GLEAM	LH	daily	all	0.25 °	global	Martens et al. (2017)
WECANN	LH	monthly	2015	1 °	global	Alemohammad et al. (2017)
	SH					
CERES	LW	daily	all	1 °	global	Wielicki et al. (1996), DOI
	SW					

* summers covered (2003/2006/2015)
** Central European domain seen by geostationary satellite, see e.g. Figure 12

small-scale differences between models and evaluation datasets can be observed and analysed. Satellite-derived land surface temperature measures T_{rad} as temperature resulting from the surface energy balance directly at the land surface (Trigo et al., 2015). It can be interpreted to be the same as the radiative temperature T_{rad} (additionally T_v in CLM) in both LSMs.

Note the different nature of missing data in SLST: While the other datasets report missing values where the earth's surface is covered by oceans or measurements failed, SLST additionally reports any value missing that is covered by clouds at the time of image capture. This results in a pattern of missing values that is different for each hourly image, making it necessary to weight daily or monthly averages by the number of valid values in this time frame for each pixel (see Table 3).

The GLEAM dataset (see Martens et al., 2017) is a model estimating evapotranspiration from remote sensing products on daily basis and a resolution of 0.25°. It provides an evapotranspiration estimate in $mm\ day^{-1}$, which is converted to Wm^{-2} with the following formula:

$$LH\ [Wm^{-2}] = -LH\ [mm\ day^{-1}]\ L_w \rho_w \frac{1}{3.6 \times 10e6 * 24} \tag{2}$$

where L_w is the latent heat of vaporisation of water in $[J\ kg^{-1}]$ and ρ_w is the density of water in $[g\ m^{-3}]$. The minus converts the dataset to the COSMO sign convention.

Additionally, we use the WECANN estimates for latent heat and sensible heat (see Alemohammad et al., 2017) derived from employing an artificial neural network on available meteorological variables, especially solar-induced fluorescence. It provides monthly estimates of the surface fluxes on a 1° resolution.

An observational ground heat flux estimate is calculated by using Equation 1.1 combining several evaluation products, once with latent heat from GLEAM and once with latent heat from WECANN:

$$G_{obs,1} = LW_{net,CERES} + SW_{net,CERES} - H_{net,WECANN} - LE_{net,GLEAM} \tag{3}$$

$$G_{obs,2} = LW_{net,CERES} + SW_{net,CERES} - H_{net,WECANN} - LE_{net,WECANN} \tag{4}$$

with all fluxes in Wm^{-2} and defined positive towards the surface. Note since ground heat flux is calculated as a residual, it also includes heat storage into vegetation. For comparison with standard procedures performed at DWD to verify COSMO-TERRA performance, we additionally compare modelled ground heat fluxes to the meteorological station in Lindenberg, Germany (see COSMO webpage).

Finally, satellite-derived observations on net shortwave and longwave radiation at the ground are obtained from the CERES dataset (see Wielicki et al., 1996).

Even though soil moisture plays a crucial role in land surface processes influencing local weather (see Section 1.1), observational data of soil moisture in sufficient coverage and quality is lacking (Seneviratne et al., 2010). This problem is exacerbated by the high spatial heterogenity of soil moisture (Seneviratne et al., 2010), hampering spatial interpolation of cost-intensive point measurements in the ground. Soil moisture satellite products have sufficient spatial coverage, but with microwave sensors, soil penetration is low and does usually not reach root zone. Gravity based methods currently have coarse resolution and cannot distinguish between water from soil, groundwater, lakes, snow or vegetation (Seneviratne et al., 2010). Since in this study gridded observations of high spatial resolution are necessary, we do not include an observational soil moisture dataset in our analysis. However, we can compare root zone soil moisture of the LSMs.

2.4 Analysis Methods

When assessing the performance of LSMs, there are three different principle routes. Following the framework of Best et al. (2015), we distinguish between comparison, evaluation and benchmarking.

Comparing different models allows to identify advantageous features of one model over another. In the light of uncertainty on how to parameterise an individual process best, it allows ranking of parameterisations and feature implementations and gives insight where model improvement is possible (Best et al., 2015). Errors systematic among models can be defined as development priorities. However, by comparing models and subsequently enhancing them, they get more alike, but not necessarily more like observations (Best et al., 2015). Furthermore, comprehensive models can not naturally bridge the gap between understanding a simulation process and understanding the physical process that is approximated with this simulation (Held, 2005).

Evaluating models is comparing models against available observations. This is typically done especially with NWP models, not only for long-term development, but also for real-time adjustment of of current forecasts to observations (data assimilation). However, a parametrisation of a certain process more aligned with observations does not necessarily lead to a more accurate representation of the process (Knutti et al., 2013). Furthermore, especially with the heavy parameterised processes on the land surface, observationally available variables are difficult to be matched with model variables, since definitions may vary. For example, the land surface temperature in the SLST dataset is the temperature "seen" by the satellite of the land surface. It is debatable whether this temperature is better compared with the radiative temperature T_{rad} defined below vegetation in CLM, since it directly results from the radiation budget, or T_v. Where vegetation is present in CLM, T_{rad} is covered by vegetation and T_v should, in principle, be the temperature "seen" by the satellite. Another example where evaluating models is difficult is the soil moisture of the root zone, crucial for land surface atmosphere coupling.

Benchmarking is a third route, where an a priori performance expectation is defined and the models ability to beat it is assessed. For example, the degree of utilizing available information of the model can be measured with a simple statistical model (Best et al., 2015). The underlying argument is as follows: While a benchmark model, for example a linear regression on selected variables, has no information about processes involved and uses little input data, a LSM has an exhaustive degree of information on vegetation state, soil state as well as water and energy budgets. This additional information, if assiduously utilized, should enable the LSM to outperform a statistical model. Especially in the context of hot summers where land surface coupling is high and soil moisture legacy from spring becomes important, LSMs should profit from their additional knowledge of the state of the land surface. It is important to note that, in this context, the statistical benchmark model should not be seen as a competitor or possible replacement of the LSM. Although fine-tuned statistical models are able to beat models of land surface, a parameterisation closely aligned to the physical processes observed are better at representing extreme cases (see Best et al., 2015) and have

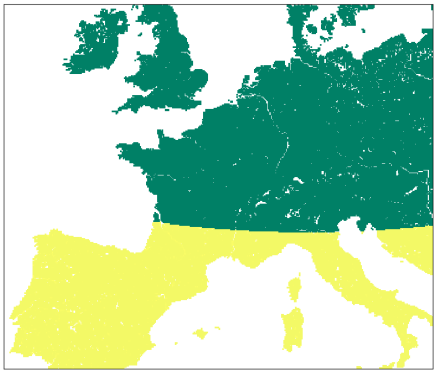


Figure 2: Division of the domain along the 45° latitude. The subdomains are in the following called Northern European and Southern European domain.

more room for improvement in future model developments. The benchmark model thus solely measures information utilization of the LSMs.

2.4.1 Evaluation and Comparison

Evaluation datasets were regridded on COSMO output using bilinear interpolation, where a point on the regridded evaluation dataset is only valid if none of the four input points is masked. Both model and evaluation datasets were additionally masked where less than 95% of the model pixel was land in either TERRA or CLM land cover input fields. In agreement with COSMO convention, all fluxes towards the surface are defined positive. The domain is divided along the 45° latitude into a Northern European domain and a Southern European domain (see Figure 2).

Aggregation to daily averages is weighted by the number of valid data points considered, e.g. for the SLST dataset, each daily mean is weighted by the number of hourly measurements it consists of. Monthly aggregated data is first aggregated daily, where applicable, and averages are also weighted by the number of valid (sub-daily) data points. Correct aggregation is ensured by calculating the weighted average on each aggregation step, which is constant throughout the aggregation steps when the weights are correct.

We apply a suite of statistical metrics (see Table 3) to evaluate model forecast with observations for all variables listed above (Section 2.3). This set of metrics informs over bias, error magnitude, correlation and shared variability of models with evaluation datasets and gives a broad overview over model performance for each variable. The metrics are weighted for number of valid measurements in the considered time period. For an overview of the biases of the model, MBE is calculated for monthly maps in summer 2015. Additionally, the representation of heat extremes in both models is estimated by applying a set of absolut heat indices (see Table 3). Absolute indices ensure comparability between models and observations and between different regions. Also, due

to the limited model run time, no relative extreme measures (e.g. comparing to a 30-year mean) can be used.

Especially for Weather Services, correctly forecasting dangerous conditions for humans is important. Thus an additional, comprehensive set of human comfort indices is applied on both models to examine the implications of the reported model bias on exposure of humans to dangerous heat events. These human comfort indices estimate apparent temperature by taking into account not only temperature, but also relative humidity and wind speed to determine temperatures felt by humans. The set consists of apparent temperature (AT), HUMIDEX and HI from Buzan et al. (2015), AT105F from Fischer and Schär (2010) and the Flanders Heat Stress Index from Wouters et al. (2017). The thresholds for discomfort and danger exposed to humans differ between the indices, but have been set so that indices exceed their limits for ”dangerous” conditions.

2.4.2 Benchmark Experiment

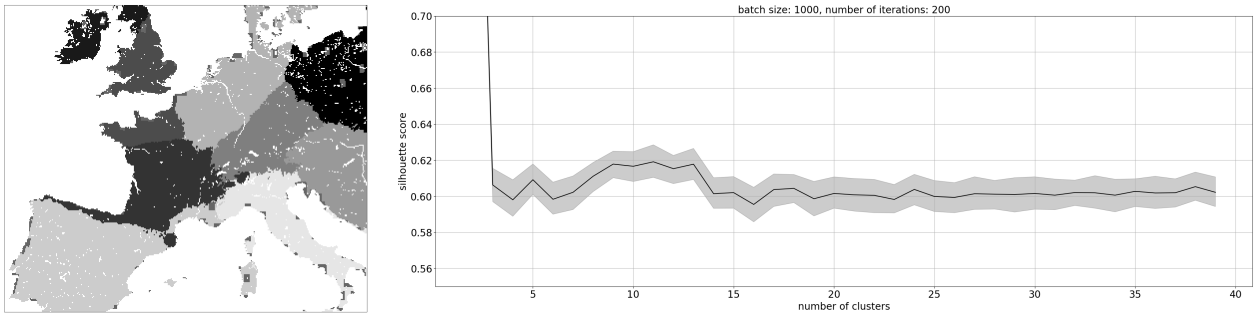
Best et al. (2015) and Abramowitz et al. (2008) showed that LSMs are unable to outperform well calibrated simple statistical models in scores describing the mean and the standard deviation of the distribution. However, there is hope for the extremes of the distribution (Best et al., 2015). To benchmark both models and analyse their capabilities in utilising information, a simple statistical model is set up to estimate latent heat flux from net shortwave radiation and precipitation provided by COSMO-TERRA. This latent heat flux is subsequently compared to the estimates of both models.

We employ precipitation and net shortwave radiation since they are the major contributing atmospheric factor to evapotranspiration. Incoming shortwave radiation is unfortunately unavailable in the current runs, therefore we use net shortwave radiation instead. Note that this already includes the shortwave radiation emitted by the land surface, and is therefore a small information advantage to the statistical model. We found the difference of using COSMO-TERRA vs COSMO-CLM precipitation and shortwave radiation minor and took the COSMO-TERRA variables as an arbitrary choice. For training the model, we use latent heat estimations from GLEAM because of the higher temporal resolution compared to WECANN. The model is trained on the summer 2006 and latent heat is estimated for the summers 2003 and 2015.

A K-means-algorithm divides the domain in 11 clusters according to their similarity regarding precipitation and shortwave radiation (Figure 3a). The algorithm minimizes within-cluster sum-of-squares, for a given number of clusters, by iteratively setting centroids of the clusters in the parameter space and moving them towards lower inertia (MacQueen, 1967). For each cluster, an individual linear ridge regression is trained, where the regularization parameter is found via cross-validation. The number of clusters is determined by calculating the silhouette score for each

Table 3: Statistical metrics applied to evaluate forecast of models with observations, where m refers to model data, o to observations and w to their respective weights. Summation over N can happen temporal, spatial or over both dimensions. BV is the best possible value for each metric.

Name	BV	Formula	Notes
Root Mean Square Error	0	$RMSE = \sqrt{\frac{\sum_{i=1}^{N} w\ (m_i - o_i)^2}{N}}$	
Mean Bias Error	0	$MBE = \overline{m_w} - \overline{o_w}$	where $\overline{m_w}$ and $\overline{o_w}$ are the weighted averages over N of model and observations, respectively
Mean Absolute Error	0	$MAE = \frac{1}{N} \sum_{i=1}^{N} w\ \lvert m_i - o_i \rvert$	
Standard deviation	—	$s_x = \sqrt{\frac{1}{N} \sum_{i=1}^{N} w\ (x_i - \overline{x})^2}$	with x as $o$bservation or $m$odel
Pearson coefficient	1	$R = \frac{\sum_{i=1}^{N} w\ (m_i - \bar{m})(o_i - \bar{o})}{N s_m s_o}$	
Regression slope	1	$\beta = \frac{s_o}{s_m} R$	
Determination Coefficient	1	R^2	
Summer Days	—	No. of days $T_{2m,max} > 25°C$	
Hot Days	—	No. of days $T_{2m,max} > 30°C$	
Tropical Nights	—	No. of days $T_{2m,min} \geq 20°C$	
CHT	—	No. of days $T_{2m,max} > 25°C$ and $T_{2m,min} > 20°C$	adapted from Fischer and Schär (2010)



(a) clusters (b) silhouette scores

Figure 3: Setup of the statistical benchmark model: (a) the domain divided into 11 clusters of similar points by a K-means-algorithm and (b) the calculated silhouette score of each number of cluster. In (b), apart from clustersize 2 which divides the domain into ocean and land, clustersize 11 is a local maximum and therefore used for the statistical model. The shading is the area between the 0.25-percentile and the 0.75-percentile of the silhouette score estimated 50 times with a batch size of 1000.

number of clusters and taking the maximum score between three and 40 clusters. We settle on this range since the global optimal silhoutte score is achieved for cluster size two, dividing the domain into land and ocean, and cluster sizes above 40 are increasingly computationally expensive.

3 Results

The model output is evaluated with gridded evaluation datasets as described in Section 2.3 where available, otherwise a comparison of the two models is performed. The results are structured as follows: First, the flux partition of both models at the surface in the summers is examined. Second, the temperature biases are characterised and the difference between the two temperature estimates is shown. An overview over model performance regarding mean and extreme indices in all included variables is then given and finally, the implications for humans from these results are shown by examining human comfort indices results of both models.

3.1 Surface Energy Balance

The absorbed energy at the surface in the LSMs is partitioned into longwave radiation, sensible heat flux, latent heat flux and ground heat flux. This partition is sensitive to LSM structure and parameterisation and differs among models. In the following, we examine the surface energy partition of both models and state differences between models compared to the evaluation datasets.

3.1.1 Latent and Sensible Heat

COSMO-TERRA overestimates sensible heat flux throughout the domain except its eastern boundaries and the Alps. In 2015, the overestimation happens most severely in July (Figure 4a) in Western Europe, where biases are more than three times larger than in June for latent and sensible heat, taking values larger than 20 Wm^{-2}. This is also the place and time where the major heat event of 2015 took place (Sippel et al., 2016). The overestimation of sensible heat follows the spatial pattern of the overestimation of latent heat, i.e. where too much sensible heat is released at the same time too little latent heat produced.

COSMO-CLM shows the same sign of bias for latent and sensible heat, but overall lower magnitudes. The overall mean bias is less than 6 Wm^{-2}. Additonally, flux estimation errors do not rise in July in COSMO-CLM.

GLEAM and WECANN overall agree regarding sign, magnitude and pattern of the bias for latent heat (Figure 4c, e and d, f), giving more confidence to the assessment. They disagree mostly in Central Europe of COSMO-CLM runs, where underestimation of latent heat is less in GLEAM than in WECANN. Errors in GLEAM are smaller than in WECANN, but RMSEs are mostly similar, suggesting more canceling of biases spatially in GLEAM. We attribute this higher spatial variability of biases in GLEAM to its higher spatial resolution.

The magnitude of the difference in COSMO-CLM and COSMO-TERRA surface energy balance (Figure 5) is shown in the evaporative fraction (EF) throughout the summer. The average evaporative fraction gives the difference in partitioning the available energy in the three summers between models. The evaporative fraction EF is calculated as:

$$EF = \frac{LH}{LH + SH} \tag{5}$$

The two available observational estimates show similar evaporative fraction throughout summer 2015: EF is around 50%. EF in COSMO-CLM closely aligns to observations for 2015 in terms of magnitude, and shows similar evolution for the two other summers, where no observations are available. In contrast, COSMO-TERRA evaporative fraction distinctively departs from observations at the beginning of July in all summers in the Southern European domain and drops to around 40%, where it stays throughout summer. This drop is most pronounced in 2015 where the major heat wave was recorded for the majority of the domain (Sippel et al., 2016).

This drop in latent heat production is not accompanied by decreasing soil moisture of the top 10 cm soil layers in TERRA (Figure 6). Soil moisture anomaly in summer is similar among models (Figure 6), but absolute values differ. COSMO-TERRA is constantly drier than COSMO-CLM in

the top 10 cm of soil, where the majority of the roots is located in TERRA.

In summary, COSMO-CLM has smaller flux at the surface biases of the same direction. COSMO-TERRA flux partition at the surface departs from both COSMO-CLM and observations at the beginning of summer and follows a different pathway throughout summer. Both latent heat observations agree well, while WECANN is unfortunately only available in 2015, stripping the opportunity to examine evaporative fraction for other summers than 2015. Coastal regions and the Alps show a different behaviour than the spatial average (see Section 4.5).

The higher the plant cover (expressed as LAI) in COSMO-TERRA, the larger is the error in the latent heat estimation (Figure 7a). A similarly strong correlation is not visible for COSMO-CLM (Figure 7b) or other combinations of variables and vegetation properties in both COSMO-CLM and COSMO-TERRA (not shown).

3.1.2 Ground Heat Flux

Ground heat flux is considerably underestimated in both models (Figure 4g and h), by on average around 30 Wm^{-2} in COSMO-CLM and around 35 Wm^{-2}, while absolute fluxes in both models are only around 15 Wm^{-2}. The underestimation is most severe in COSMO-TERRA in Southern Spain. The underestimated heat storage capacity in the ground in both models intensifies towards summer (Figure 8a and b): ground heat flux is negative or close to zero throughout January and starts to get positive in February. The seasonal cycle of ground heat flux is larger in COSMO-CLM than in COSMO-TERRA, leading to lower bias compared to the evaluation data. However, both models are more similar to each other than to the observations. The single years show departures from this behaviour in single outlier events, which are most common in January.

To investigate this striking difference between models and observational estimate, both models are additionally compared ground heat flux in 5 cm depth at the meteorological station in Lindenberg, Germany. For that purpose, the pixel in both models which contains the measurement station is extracted and compared to the observations at the station (Figure 8c). In contrast to the gridded ground heat flux estimate, ground heat flux at Lindenberg station has a smaller magnitude than in the models on the seasonal time scale (Figure 8c).

Note that the vegetation at Lindenberg is grassland, while the model pixel extracted is only 53% grassland in TERRA and 17% grassland (43% crop) in CLM, hampering comparison between the observations and modeled fluxes. Note furthermore that the pixel containing the Lindenberg station is usually masked in this analysis, because there are lakes in the vicinity of the station that decrease the land fraction of the pixel to a value lower than 0.95.

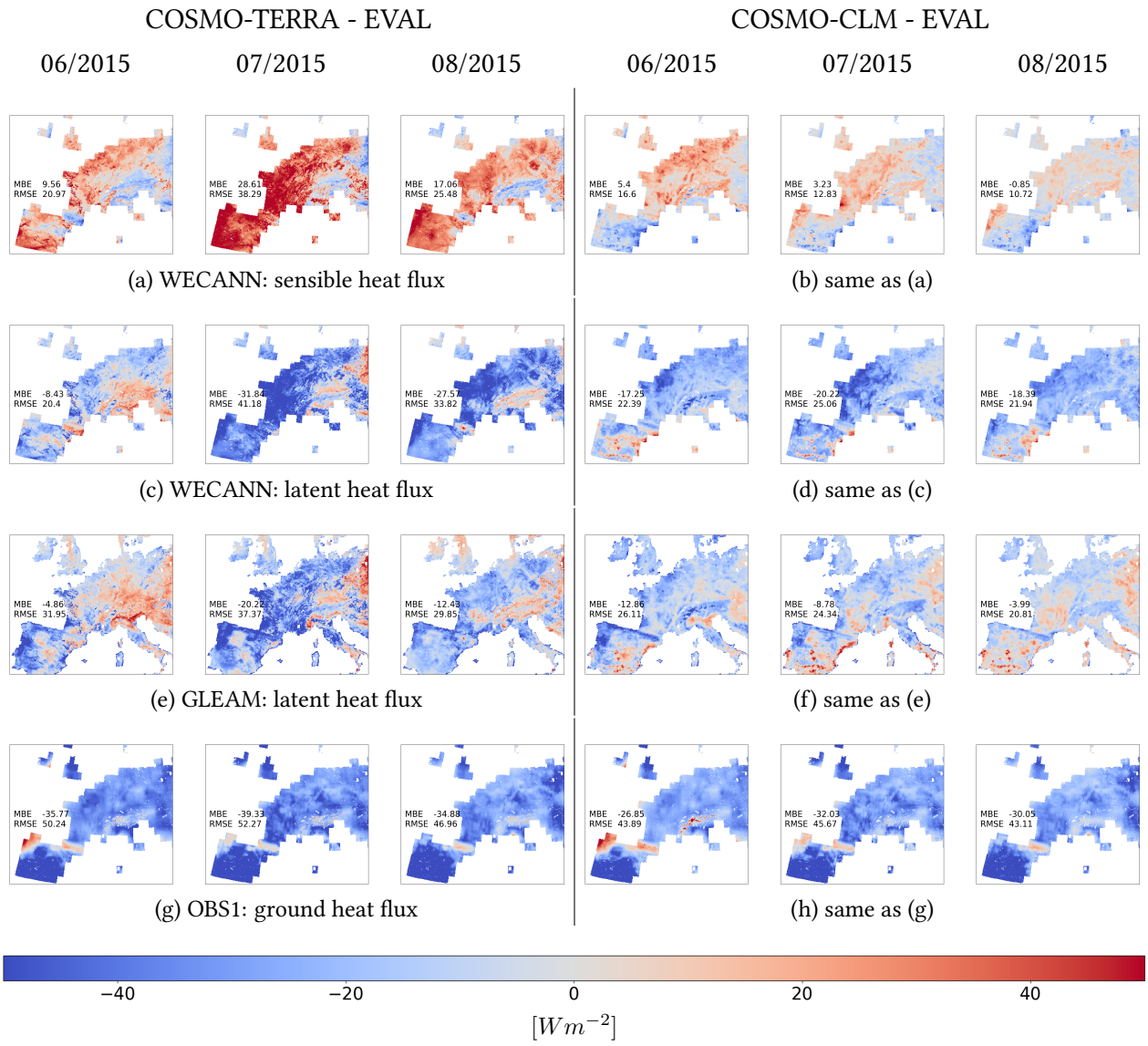
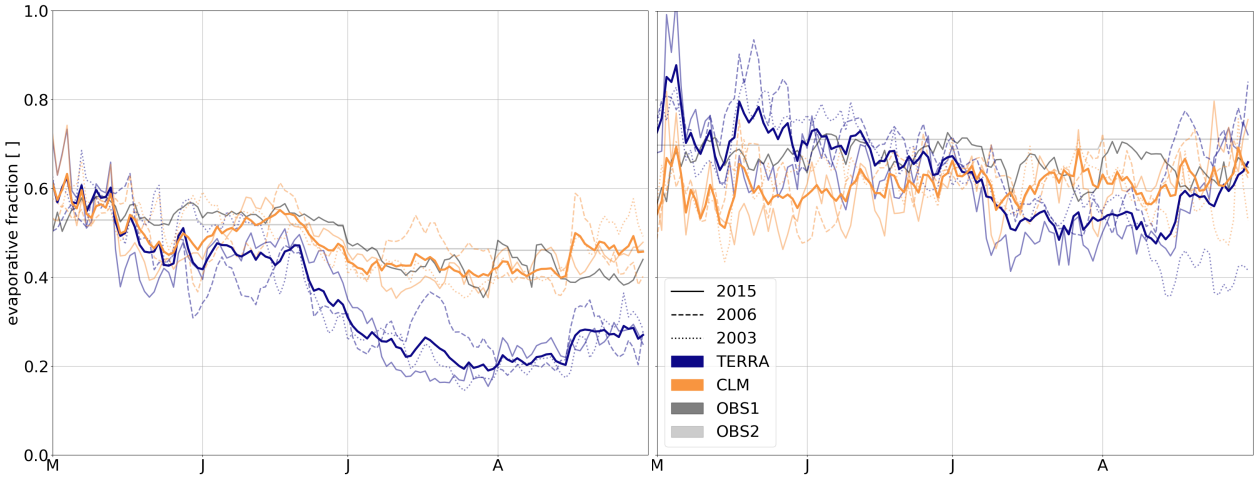
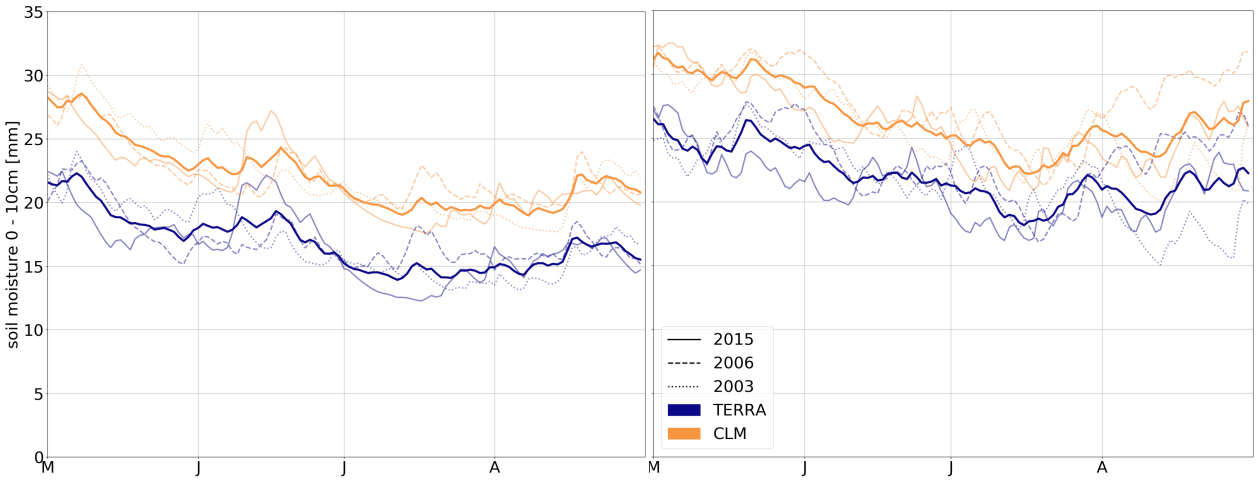


Figure 4: Monthly mean bias (model minus observations) in daily mean (a, b, c, d) sensible heatflux (e, f) latent heat flux and (g, h) ground heat flux for COSMO-TERRA (a, c, e, g) and COSMO-CLM (b, d, f, h) as compared to the WECANN dataset (a, b, c, d), the GLEAM dataset (e, f) and the observational ground heat flux derived from WECANN sensible heat and GLEAM latent heat estimates (g, h) for JJA 2015. Positive values indicate overestimation of the flux in models. RMSE and MBE for each month are printed on the maps. Data where no measurements are available in observations are excluded.



(a) Southern European domain (b) Southern European domain

Figure 5: Mean daily evaporative fraction in COSMO-TERRA, COSMO-CLM and observations averaged over (a) the Southern and (b) the Northern European domain for JJA in 2003, 2006, 2015 and the average of the three summers. Observational evaporative fraction has two estimates, one with latent heat from GLEAM and sensible heat from WECANN (OBS1) and a second one with both variables from WECANN (OBS2) (see Equations 3 and 4). Note that the observational evaporative fraction can only be computed for 2015, since the WECANN dataset does not stretch back to the other summers. CHANGE TITLE



(a) Southern European domain (b) Northern European domain

Figure 6: Mean daily root zone soil moisture (0 - 10 cm depth) COSMO-TERRA and COSMO-CLM averaged over the (a) Southern and (b) Northern European domain for JJA in 2003, 2006, 2015 and the average of the three summers.

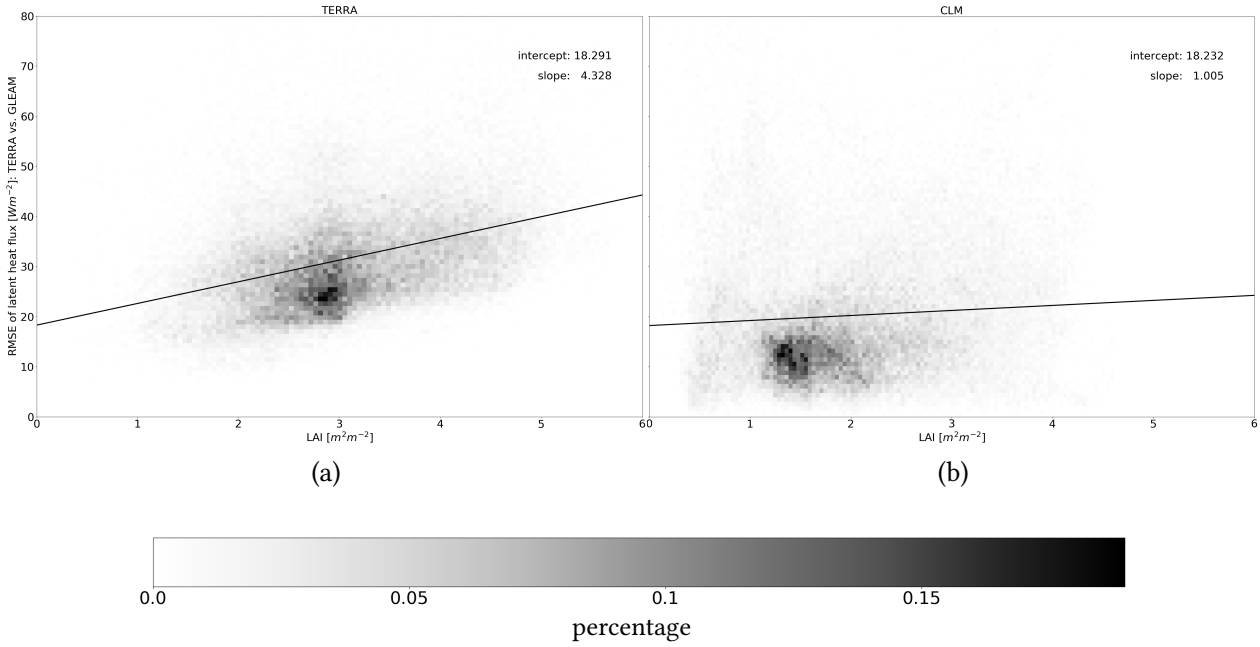


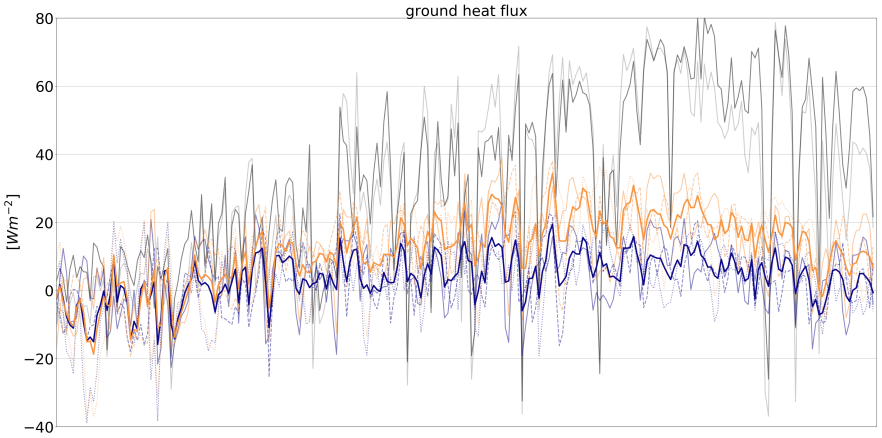
Figure 7: Dependency of the RMSE in the latent heat estimation of (a) COSMO-TERRA and (b) COSMO-CLM on the mean LAI. Daily values are integrated over the whole domain and JJA 2015.

3.1.3 Mean Diurnal Cycle of Fluxes

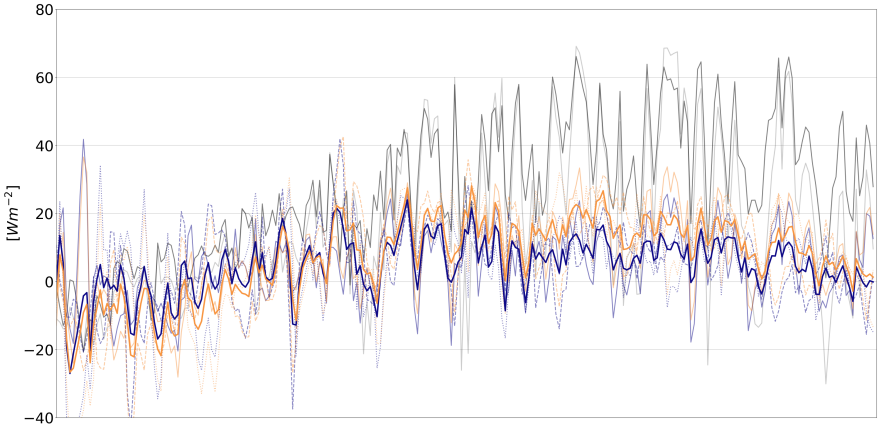
Net shortwave radiation in both models is comparable (Supplementary Figure A8a, b and Figure 9), with COSMO-TERRA having on average higher shortwave radiation throughout Europe. Longwave radiation is overestimated compared to observations (Supplementary Figure A8c, d) but similar among models (Figure 9). The partition into the surface fluxes (sensible heat, latent heat and ground heat flux) on sub-daily time scales is where the main differences between the two models start to show.

In the Southern European domain, overall larger flux magnitudes lead to larger differences. Additionally, here the impact of the difference in the evaporative fraction is higher: COSMO-TERRA produces more sensible heat, and less latent heat than COSMO-CLM. This behaviour is strikingly consistent between years, more than between models (Figure 9). Over the whole domain, the residual ground heat flux is larger in COSMO-CLM: the LSM stores more heat in the ground during day and releases the majority in a distinctive event in the early afternoon (Supplementary Figure A10).

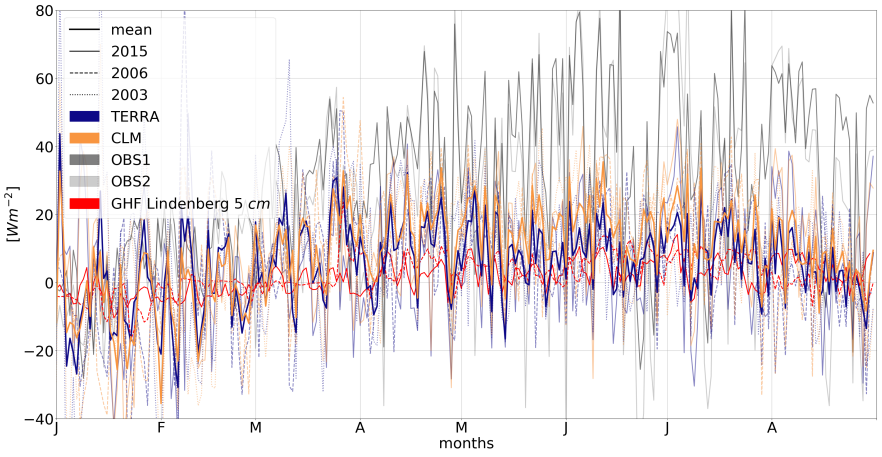
Diurnal cycles of both models at Lindenberg are comparable to mean diurnal cycles over the whole domain in terms of shape and magnitude (Figure 10). Sensible heat flux in models is overstimated and latent heat is underestimated compared to Lindenberg (Figure 10b, c). This is in agreement with results for our gridded datasets, but in contrast to the findings from Schulz and Vogel (2017).



(a) Southern European domain

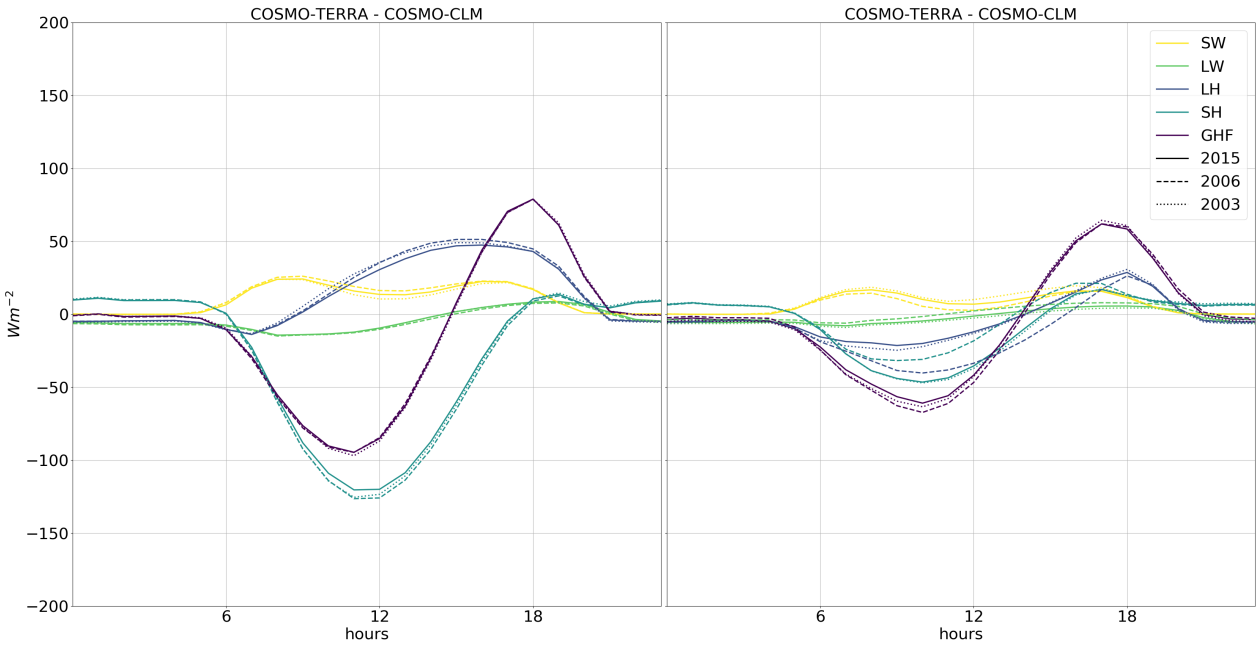


(b) Northern European domain



(c) Lindenberg

Figure 8: Domain averaged ground heat flux over (a) the Southern European, (b) the Northern European domain and (c) of the model point at Lindenberg as daily averages for the first half of the three years. Note that an observational ground heat flux estimate is only available for 2015 and a full seasonal cycle is not available since model runs do not cover all months.

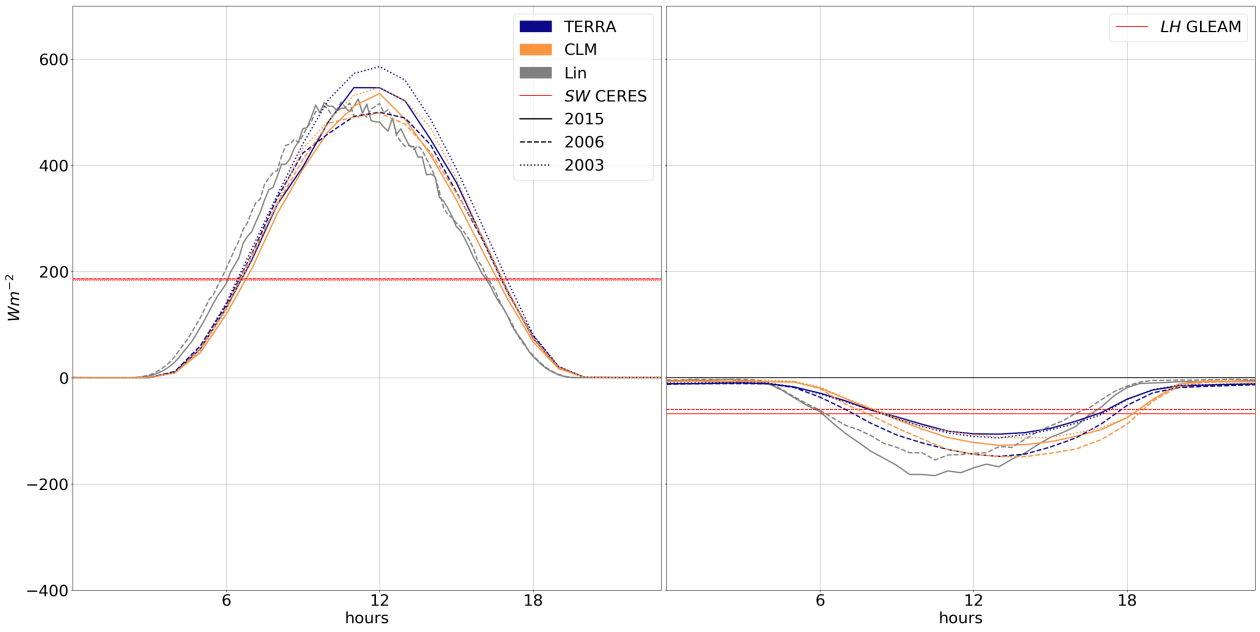


(a) Southern European domain (b) Northern European domain

Figure 9: Difference of mean diurnal cycle of net shortwave radiation (SW), net longwave radiation (LW), net latent heat (LH) and net sensible heat (SH) at the surface between COSMO-TERRA and COSMO-CLM (COSMO-TERRA - COSMO-CLM) averaged over the (a) Southern European and (b) Northern European domain for JJA in 2003, 2006 and 2015. Note that in agreement with the COSMO convention, fluxes towards the surface are defined positive. Positive values thus denote *more positive* flux towards the surface in COSMO-TERRA, while negative values denote more positive flux towards the surface in COSMO-CLM. For mean diurnal cycle of both models separately see Supplementary Figure A10.

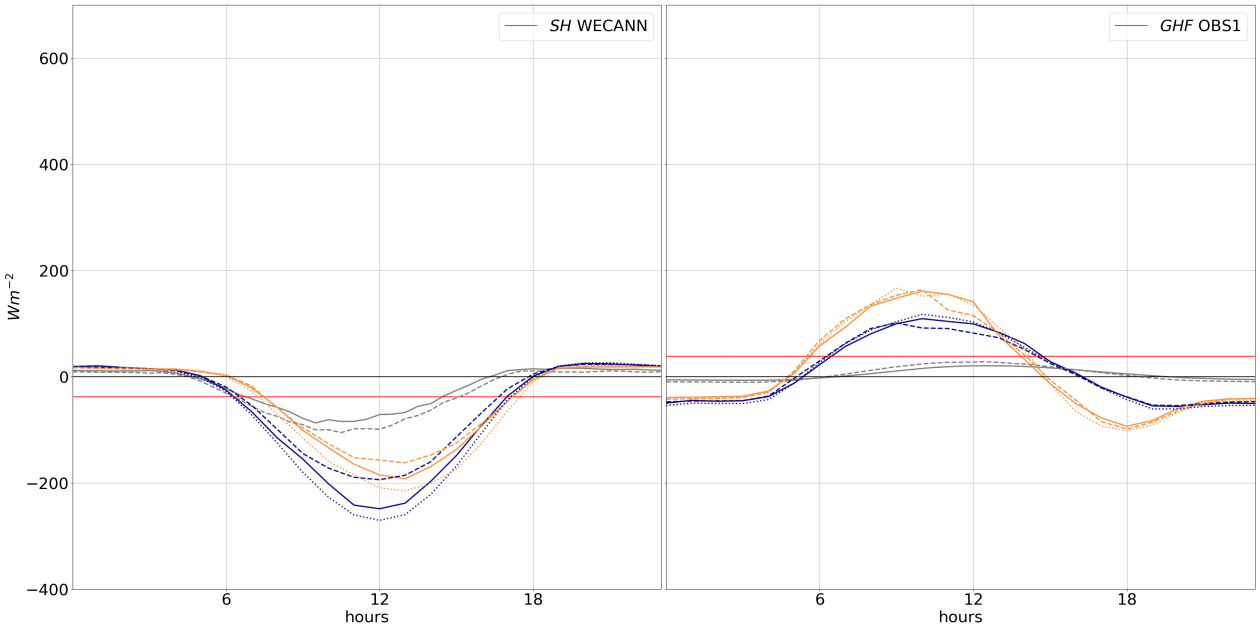
Lindenberg shortwave and longwave radiation are comparable to model estimates (Figure 10a and Supplementary Figure A13, respectively).

Ground heat flux in both models exhibits a strong diurnal cycle of around 250 Wm^{-2} in COSMO-CLM and 150 Wm^{-2} in COSMO-TERRA, while the diurnal cycle at Lindenberg is around 50 Wm^{-2} at maximum (Figure 10d). Ground heat flux at Lindenberg reaches its daily maximum two to three hours later than in the models and does not show a global minimum in the afternoon as COSMO-CLM does. However, we expect ground heat flux to be dampened and delayed at the depths measured at Lindenberg, so comparing the magnitude of both fluxes is challenging. A scaling of the ground heat fluxes of both models to the depth of observations in Lindenberg was unfortunately outside of the scope of this master thesis.



(a) Shortwave radiation

(b) Latent heat flux



(c) Sensible heat flux

(d) Ground heat flux

Figure 10: Comparison of diurnal cycle of (a) shortwave radiation, (b) latent heat flux, (c) sensible heat flux (g) ground heat flux and longwave radiation (Supplementary Figure A13) at the Lindenberg station and the two models at the associated model pixel averaged over the summer months for the three summers. The respective observational daily estimate is included (red line). Note furthermore that ground heat flux at Lindenberg is measured in 5 cm depth.

3.2 Model Temperature Bias

Temperature bias for both 2-meter temperature and radiative temperature agrees among models, but in comparison to the two available evaluation datasets bias is different.

3.2.1 2-Meter Temperature

Similarly to latent heat, sensible heat and ground heat flux, both models show a bias that is broadly similar in sign, but different in magnitude (Figure 11) when evaluated against EOBS. Observed temperature biases at 2 meters reflect the partition of the energy at the surface in both models: The warm bias in COSMO-TERRA is in line with an excess of sensible heat release at the cost of too little latent heat production.

2-Meter mean temperature in COSMO-CLM is best of all the variables and models shown in this plot, with in small scale, patchy biases and an MBE around 0.5 K. Larger biases only appear in the Alpine region. COSMO-TERRA overestimates daily 2-meter mean temperature over almost the whole domain by about 1 K, except for small regions especially in the northern coastal area of Spain, southern Italy and the higher Alpine regions.

2-Meter minimum temperature has the largest bias in both COSMO-TERRA and COSMO-CLM. It is consistently overestimated throughout the domain. COSMO-TERRA is up to 2 K too warm, while COSMO-CLM has mostly smaller biases of around 1.5 K.

2-Meter maximum temperature agrees less between models. COSMO-CLM mostly underestimates 2-meter maximum temperature, especially on the Iberian Peninsula and north of the Alps. In COSMO-TERRA, there is no uniform bias in space or time. The eastern part of the domain is underestimated, while overestimation occurs over the rest of the domain. In June TERRA is overall too cold, while in July and August it is too warm.

2-Meter diurnal cycle: Largely overestimated 2-meter minimum temperature combined with underestimated 2-meter maximum temperature in parts of the domain translates into a reduced diurnal cycle of 1-2 K of similar magnitude both models as compared to observational products (Figure 11g).

Figure 13a shows that this behaviour of large warm bias at night and smaller cold bias at day is consistent throughout the three summers, happens both in Northern and Southern Europe and is similar between models. However, the diurnal cycle in COSMO-CLM and COSMO-TERRA have shifted biases: COSMO-CLM is colder during day and night than COSMO-TERRA in the

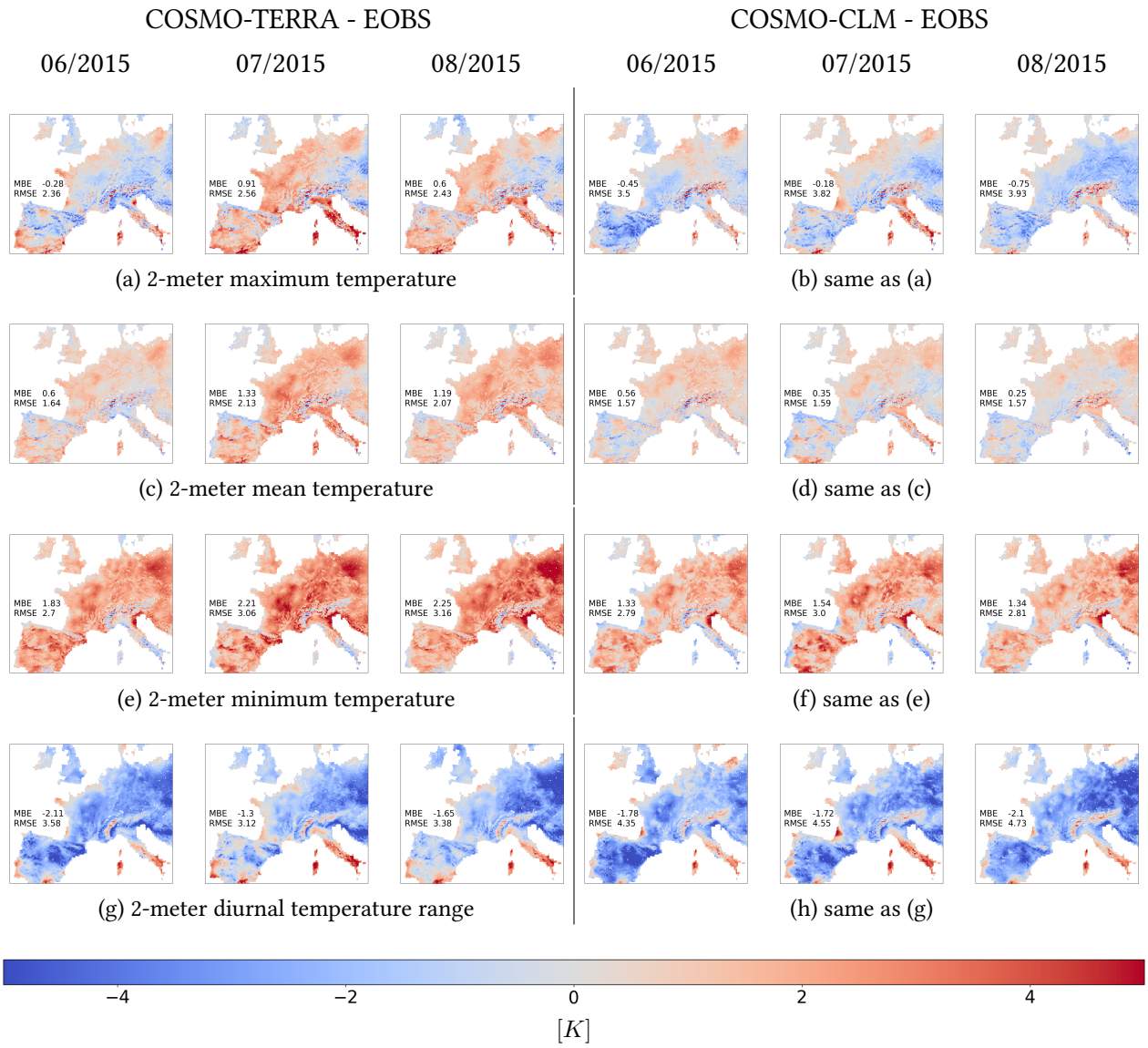


Figure 11: Monthly mean bias (model - observations) in (a, b) daily maximum 2-meter temperature, (c, d) daily mean 2-meter temperature, (e, f) daily minimum 2-meter temperature and (g, h) diurnal 2-meter temperature range for COSMO-TERRA (a, c, e, g) and COSMO-CLM (b, d, f, e) as compared to the EOBS dataset for JJA 2015. RMSE and MBE for each month are printed on the maps. Data where no measurements are available in EOBS are excluded. See Appendix for bias maps of 2006 (Supplementary Figure A3) and 2003 (Supplementary Figure A2).

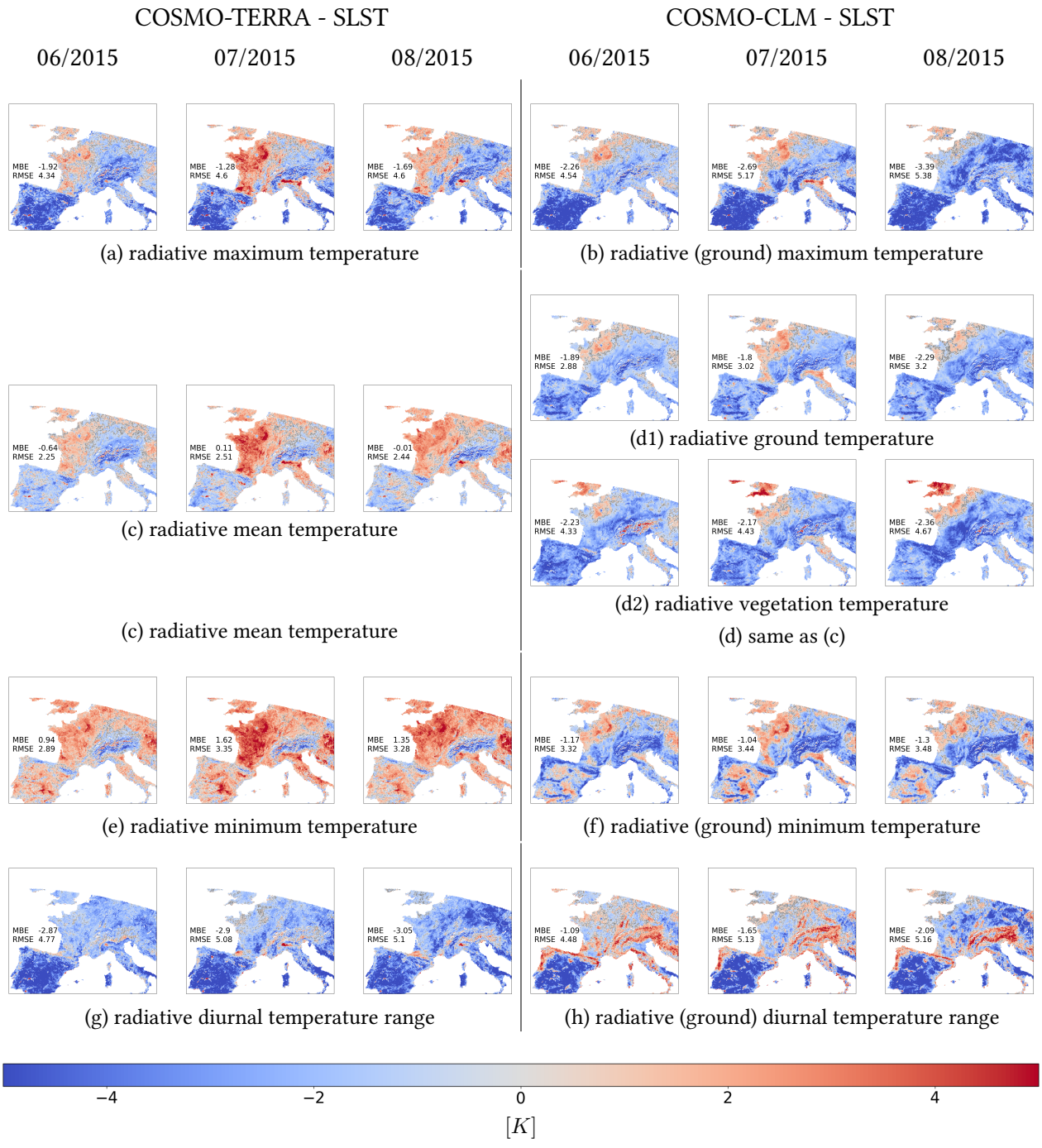
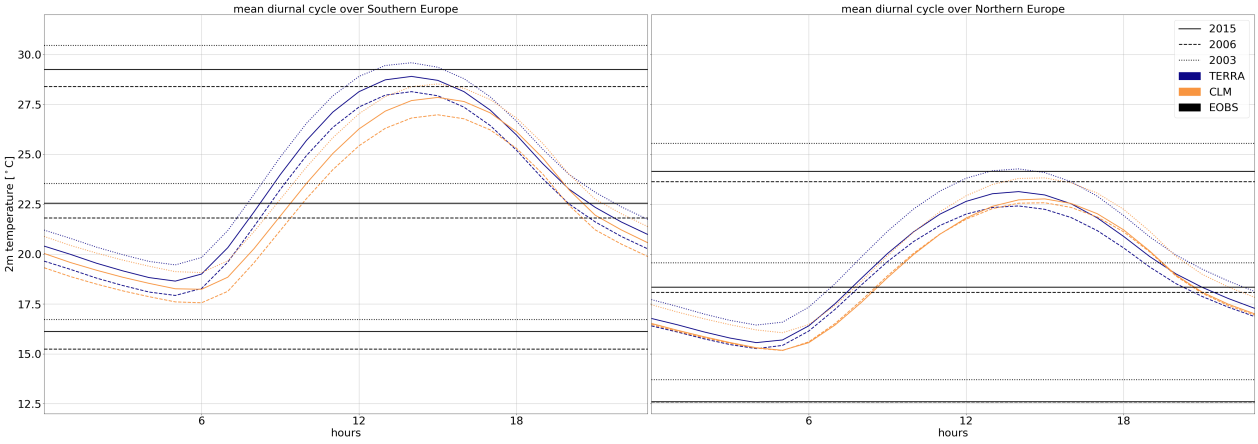
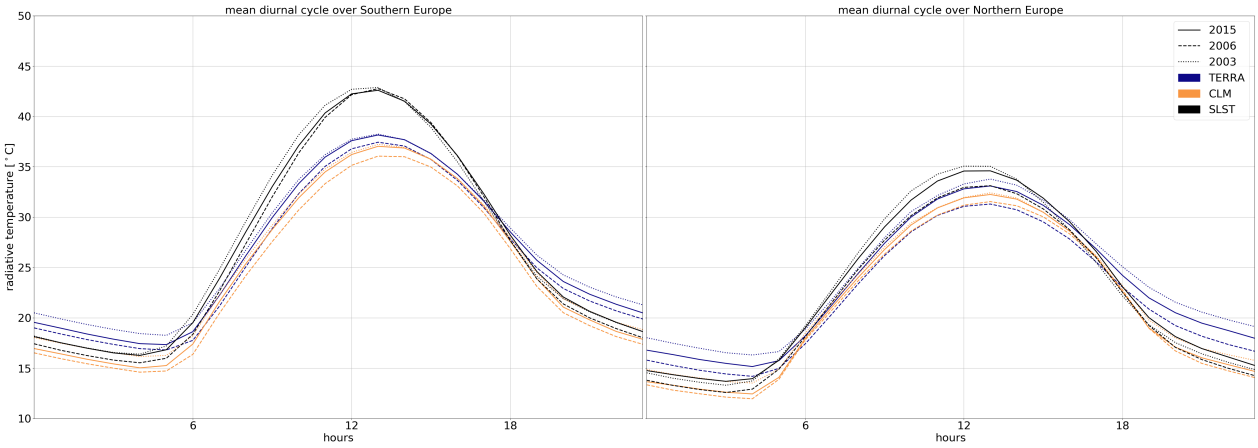


Figure 12: Monthly mean bias (model - observations) in (a, b) daily maximum radiative temperature, (c, d) daily mean radiative temperature, (e, f) daily minimum radiative temperature and (g, h) diurnal radiative temperature range for COSMO-TERRA (a, c, e, g) and COSMO-CLM (b, d, f, h) as compared to the SLST dataset for JJA 2015. For evaluation with COSMO-CLM, the CLM variable ground temperature is used in all plots except (d2), where vegetation temperature is shown additionally for comparison (see Section 2.2 for distinction). Points are stippled where the uncertainty of the satellite measurement is larger than the reported bias at this point. RMSE and MBE for each month are printed on the maps. Data where no measurements are available in SLST are excluded. Note that this especially includes all data points where cloud cover was detected by the satellite. See Appendix for bias maps of 2006 (Supplementary Figure A5) and 2003 (Supplementary Figure A4).



(a) 2-meter temperature

(b) same as a



(c) radiative temperature

(d) same as c

Figure 13: Mean diurnal cycle of (a, b) 2-meter temperature and (c, d) radiative temperature of COSMO-TERRA, COSMO-CLM and available observations averaged over the Southern European domain (a, c) and Northern European domain (b, d). Observations for 2-meter temperature are daily EOBS data, where mean, maximum and minimum temperature per day is recorded, and SLST data for radiative temperature, which provides hourly output. Note that data where no measurements are available in the observations are excluded, i.e. sampling only cloud-free data in c and d.

Northern and Southern European domain. Furthermore, the daily minimum and maximum 2-meter temperature in COSMO-CLM occur on average one hour later in both subdomains.

COSMO-CLM performs better in terms of 2-meter minimum and mean temperature, because it is colder than COSMO-TERRA, while COSMO-TERRA represents 2-meter maximum temperature better, because it is warmer than COSMO-CLM. These overall trends have departures especially in the Alps and coastal regions (see Section 4.5) but show, considering the extent and the variability in terms of land surface throughout the domain, striking spatial similarity.

3.2.2 Radiative Temperature

A second evaluation of modelled temperature was performed with the SLST dataset (Figure 12). Note the SLST dataset records radiative temperature, which is different from 2-meter temperature. The radiative temperature of the SLST dataset was compared to T_{rad} in COSMO-TERRA and to T_{rad} *and* T_v in COSMO-CLM (see Section 2.2). For T_v, only daily mean temperature output is available. Note furthermore that the satellite data only provides measurements where no cloud cover is present, i.e. the following results only sample cloud-free conditions (as seen by the satellite).

The results of the comparison with SLST for COSMO-CLM do not mirror the findings in the fluxes and for 2-meter temperature: the models do not agree in the sign of their biases. COSMO-CLM is on average too cold while COSMO-TERRA shows better mean and maximum temperatures representation because it is warmer. RMSEs of both models are of similar range for all variables, but COSMO-TERRA often outperforms COSMO-CLM in 2015.

Radiative mean temperature bias is overall positive in Western France, the British Islands, Northern Italy and parts of Hungary and overall negative on the Iberian Peninsula and in the Alps (Figure 12c). In COSMO-TERRA, mean temperature shows canceling biases between maximum and minimum temperatures and has the smallest errors among all radiative temperature variables.

Both COSMO-CLM mean temperature variables have higher RMSE than COSMO-TERRA. T_{rad} is colder than SLST radiative temperature except for parts in northern France and the British Islands, on average around 2 K. T_v shows similar patterns, with overall higher under- and overestimation, on average more than 2 K too cold.

Radiative minimum temperature: Nighttime temperature bias differs among models. COSMO-TERRA is on average 1 K too warm at night, while COSMO-CLM is 1 K too cold at night over the whole domain. Figure 13c and d shows a similar bias for Northern and Southern Europe.

Radiative maximum temperature in COSMO-CLM and in COSMO-TERRA is especially largely underestimated on the Iberian Peninsula (see also the Southern subdomain in Figure 13c). In Northern Europe (see Figure 13d), there is little warm bias at day for both models. The underestimation is larger in COSMO-CLM than in in COSMO-TERRA. RMSEs are similar in both models, indicating spatially compensating errors in COSMO-TERRA.

Radiative diurnal cycle: Underestimated daytime temperatures and overestimated nighttime temperatures in COSMO-TERRA lead to a considerable and consistent reduced diurnal cycle of up to 3 K over almost all the domain.

For COSMO-CLM, we see a mixed picture of diurnal temperature range performance: While especially in central Spain and north of the Alps the diurnal cycle is too small, coastal areas in Spain and the Alps have an overestimated diurnal temperature range.

For both models (see Figure 13b), the most prominent bias is a considerable 3-4 K cold bias in the models during day in Southern Europe.

Between summers, the models behave consistent in all discussed variables (see Figure 13). SLST provides an error estimate of the measured radiative temperature. In Figure 12 points are stippled where the uncertainty of the measurement is higher than the reported bias between model and observation. As this is only happening in regions where bias is small and changing its sign, it gives us confidence for the reported bias.

3.2.3 2-Meter Temperature vs. Radiative Temperature

Comparing results for 2-meter temperatures and radiative temperatures turns out to be challenging. Both temperature variables show a cold bias of COSMO-CLM as compared to COSMO-TERRA. This cold bias, however, is beneficial to model performance compared with EOBS but degrades model performance compared to SLST. Thus, COSMO-CLM outperforms COSMO-TERRA at 2 meters except for maximum temperatures, and COSMO-TERRA outperforms COSMO-CLM for radiative temperatures, except for minimum temperatures.

When comparing to EOBS, the biases are smaller and variables tend to have overall similar biases. When comparing to SLST the biases are larger, often cancel each other out and the spatial patterns in which this happens is consistent among variables, but not similar to the patterns in Figure 11. Higher spatial variability in radiative temperatures most likely stems from higher spatial and temporal resolution of the SLST dataset, allowing for more regional differences to be displayed. Higher temporal variability in radiative temperature on sub-daily timescales comes from the different nature of the two variables: Since temperature ranges at 2 meters are already dampened, radiative temperature exhibits a larger diurnal cycle than 2-meter temperature.

3.3 Scores for Mean and Extreme Indices

An overall verification of both models is given by applying a suite of metrics for mean and extreme indices. COSMO-CLM has smaller errors, larger correlation and similar variability in the fluxes compared to the evaluation datasets. However, this improved representation of radiative and surface fluxes does not translate into the temperatures, where a more complex picture arises (Figure 14).

Radiative and Surface Fluxes: Over all three summers, COSMO-CLM consistently has smaller errors in and larger correlation when representing fluxes (Figure 14). The only exception is sensible heat, where values in COSMO-TERRA correlate better with WECANN observations. Ground heat flux estimates and shortwave radiation estimates have the largest errors. For shortwave radiation estimates, this is expected since they also take the largest absolute values. The correlation of shortwave radiation, independent of the absolute values, indicates similar performance than in the other fluxes. COSMO-TERRA variability in fluxes is considerably higher than observed, while COSMO-CLM is closer to observations. For COSMO-CLM this especially boosts its β value. R^2 values show low proportion of variance explained in the fluxes, especially latent heat in GLEAM and ground heat flux.

Temperatures have better correlation than fluxes in both models. This is not surprising, since models are tuned to get temperatures right. In terms of error scores, COSMO-CLM outperforms COSMO-TERRA at 2 meters except for maximum temperatures, and COSMO-TERRA outperforms COSMO-CLM for radiative temperatures, except for minimum temperatures. The differences between the two models regarding minimum temperature, however, are within the range of variability between the three years. T_{rad} in COSMO-CLM is a better estimate for radiative temperature than T_v. Correlation measures paint a different picture: 2-Meter minimum temperature has better correlation in COSMO-TERRA. Differences for radiative temperature are within the variability range of the three years. All different correlation metrics show the same model preference. Overall best performance is achieved in mean temperatures. COSMO-CLM shows more similar variability compared to EOBS, but less similar variability compared to SLST. Maximum radiative temperature is more variable in observations than in models and thus gets a boost for β.

A bias map of the latent heat estimated by the benchmark is shown in Figure 15. Note that the benchmark was trained with GLEAM data from summer 2006, so the comparison is favored towards GLEAM. The tiles produced by the K-means-algorithm (see Figure 3b) can be spotted easily. The K-means-algorithm divides the domain into clusters of geographically close points. The benchmark therefore does not show physically consistent behaviour along the tile borders. However, this is not the purpose of this experiment (see 2.4.2 for further discussion).

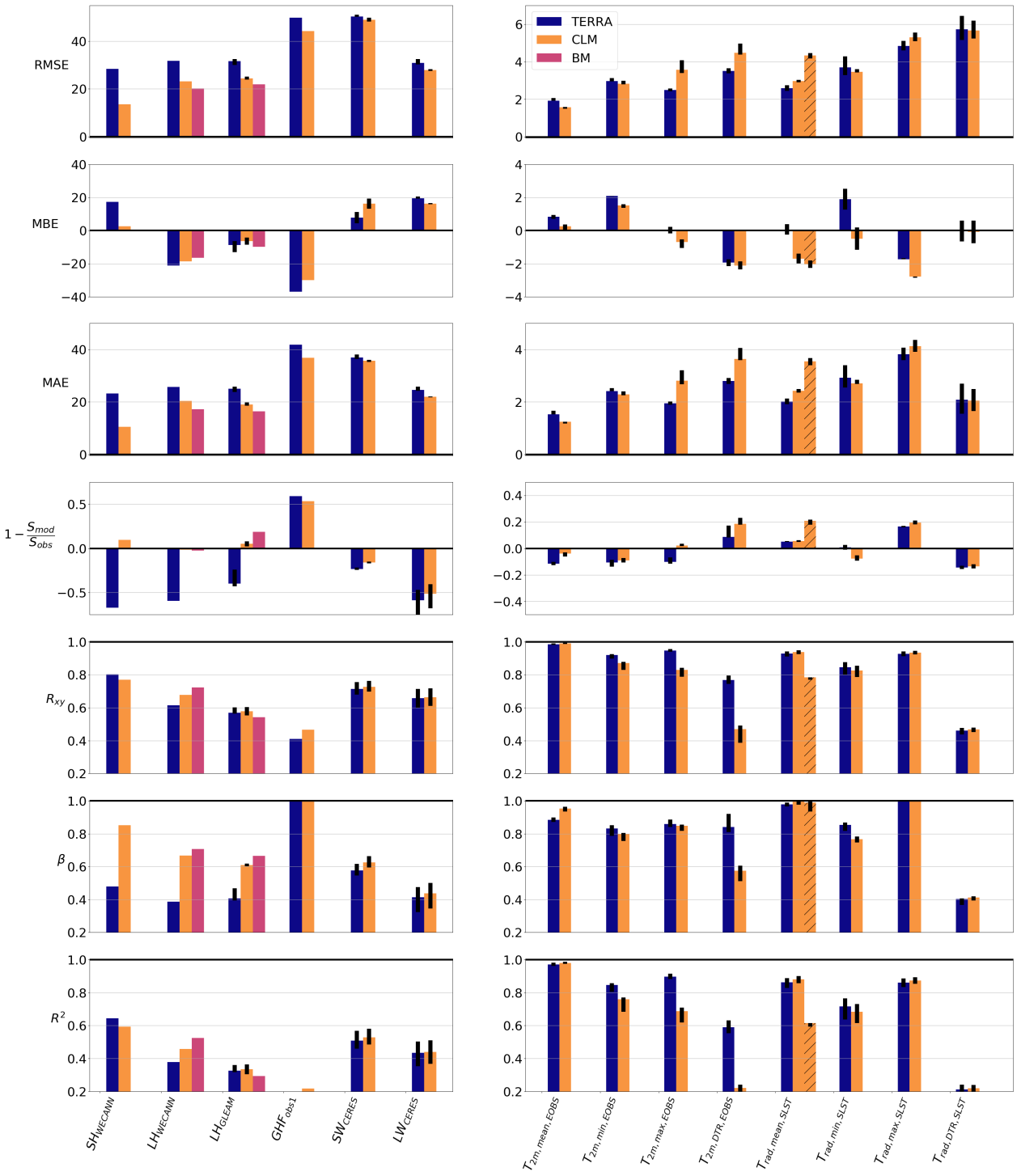
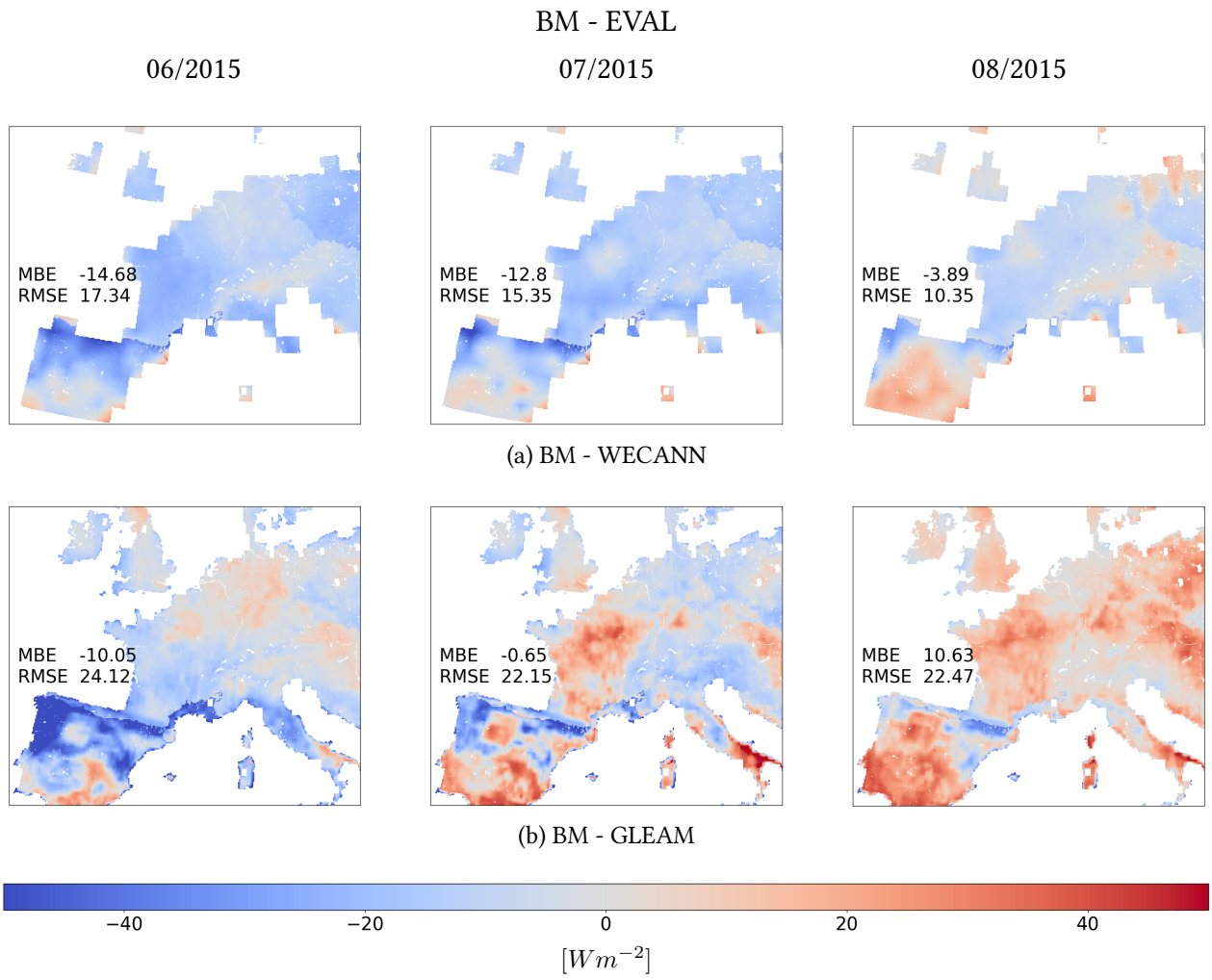


Figure 14: Performance metrics with all evaluation datasets available. Measures are calculated for daily values in JJA of 2015 and 2003. 2006 is left out since no benchmark is available in this year. For $T_{rad,mean,SLST}$, hatched bars are comparison to T_v and unhatched bars for T_{rad} in COSMO-CLM. For WECANN, only 2015 is available. Black thick lines indicate the best value of each measure. Errorbars indicate the difference between the two years. Plot is also available for all three summers (Supplementary Figure A14), and individually for the Southern (Supplementary Figure A15) and Northern European doman (Supplementary Figure A16). For similar analysis in climate mode see Supplementary Figure A17.



(a) BM - WECANN

(b) BM - GLEAM

Figure 15: Monthly mean bias (benchmark - observations) in daily mean latent heat flux as compared to the WECANN dataset (a) and the GLEAM dataset (b) for JJA 2015. RMSE and MBE for each month are printed on the maps. Data where no measurements are available in observations are excluded. For a comparison of estimated fluxes in GLEAM, the models and the benchmark see Supplementary Figure A11.

The benchmark is always better than COSMO-TERRA and COSMO-CLM only reaches similar performance when compared to GLEAM. The benchmark shows considerably lower variability than observations. When this low variability is taken into account, as for example with β, the benchmark far outperforms both models also compared to GLEAM. The benchmark suffers from the same biases than both models: it underestimates latent heat and overestimates sensible heat. In the beginning of summer it is more underestimated and starts to overestimate towards the end of summer for GLEAM and WECANN on a lesser degree. The errors are similar between months, notably, it shows no error peak in July as COSMO-TERRA does.

The benchmark outperforms both models when evaluated with WECANN and GLEAM. COSMO-TERRA has considerably larger errors and less correlation compared to the benchmark, COSMO-CLM is quite close to the benchmark.

The improved performance regarding land surface energy balance in COSMO-CLM did not translate into the temperatures, where a consistent cold bias of COSMO-CLM is hampering its performance. However, when applying heat extreme indices on both models, COSMO-CLM shows better performance.

Representation of heat extremes in both models agrees spatially with observations, but magnitudes differ (Figure 16). High occurrence of extreme events in the summers occur on the Iberian Peninsula, in low-altitude Italy and Western France. When only maximum temperatures plays into the index, COSMO-CLM underestimates extreme event occurrence, while COSMO-TERRA overestimates it. When additionally night-time temperatures are taken into account, the model performance diverges: COSMO-CLM overestimates heat extreme occurrence less than COSMO-TERRA, since it overestimates nighttime 2-meter temperatures less than COSMO-TERRA. In Summary, for indices depending solely on temperature, COSMO-CLM has better heat extreme representation than COSMO-TERRA.

3.4 Human Comfort Indices

The impact of heat extremes on humans does not only depend on temperature. Severity of heatwaves can be approximated by the apparent temperature felt by humans, also depending on humidity and wind speed. To investigate the implications for humans of the observed differences, a set of human comfort indices is applied on both models (Figure 17). These measures include, among temperature, also humidity and wind speed. The latter variables are not available in a gridded dataset for the domain, therefore only a comparison of both models, not an evaluation against observations, is possible here. Threshold exceedances are in the following called AT30 (for AT $> 30\ ^{\circ}C$), HI40 (for HI $> 40\ ^{\circ}C$) and HUMIDEX46 (for HUMIDEX > 46).

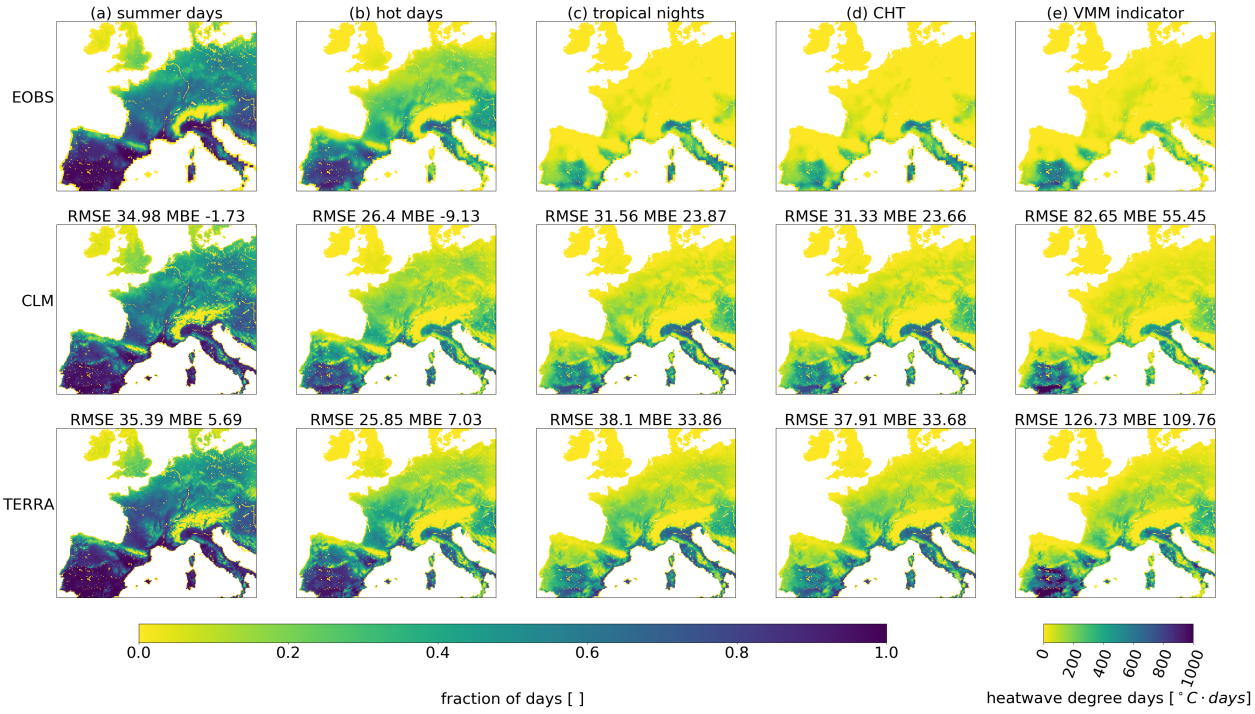


Figure 16: Occurence of heat extremes over Europe for all three summers expressed as fraction of days which classify as (a) summer days, (b) hot days, (c) tropical nights and (d) both hot day and tropical night (CHT, see Fischer and Schär, 2010) in EOBS (upper panels), COSMO-CLM (middle panels) and COSMO-TERRA (lower panels). (E) shows the Flanders heat index (from Wouters et al., 2017) for the same time period (note different unit). Titles of middle and lower panels are RMSE of modeled heat extreme incidence with observations (upper panels).

Areas often identified as extreme in the heat extreme metrics in Section 3.3 do also often cross thresholds of apparent temperature estimates. High human comfort indices are reported in the southern half of the Iberian Peninsula and throughout low-altitude Italy. COSMO-TERRA and COSMO-CLM share broadly the same patterns. Between indices, however, striking differences arise. HI40 and AT105F detect high occurrence, amplitude and duration of heat extremes in the Alpine region, a feature unknown to the other indices discussed. HUMIDEX is reporting far less, but higher extreme events than the other indices. It is unitless, which means its amplitude is not comparable to other indices. The threshold set by the authors (see Masterton, 1979), however, detects far less extreme events than the other indices for similar conditions. Amplitudes are similar among human comfort indices except AT30.

Model differences for human comfort indices are local and small-scale. A clear signal only emerges from AT30 events, which are more often, more prolonged and slightly warmer in COSMO-TERRA especially on the Iberian Peninsula and in France. HI40 and AT105F, on the contrary, show lower amplitudes in France for COSMO-TERRA. The regions of disagreement are uniform across all indices, and are at the same time the regions where high heat extremes are reported in the three summers.

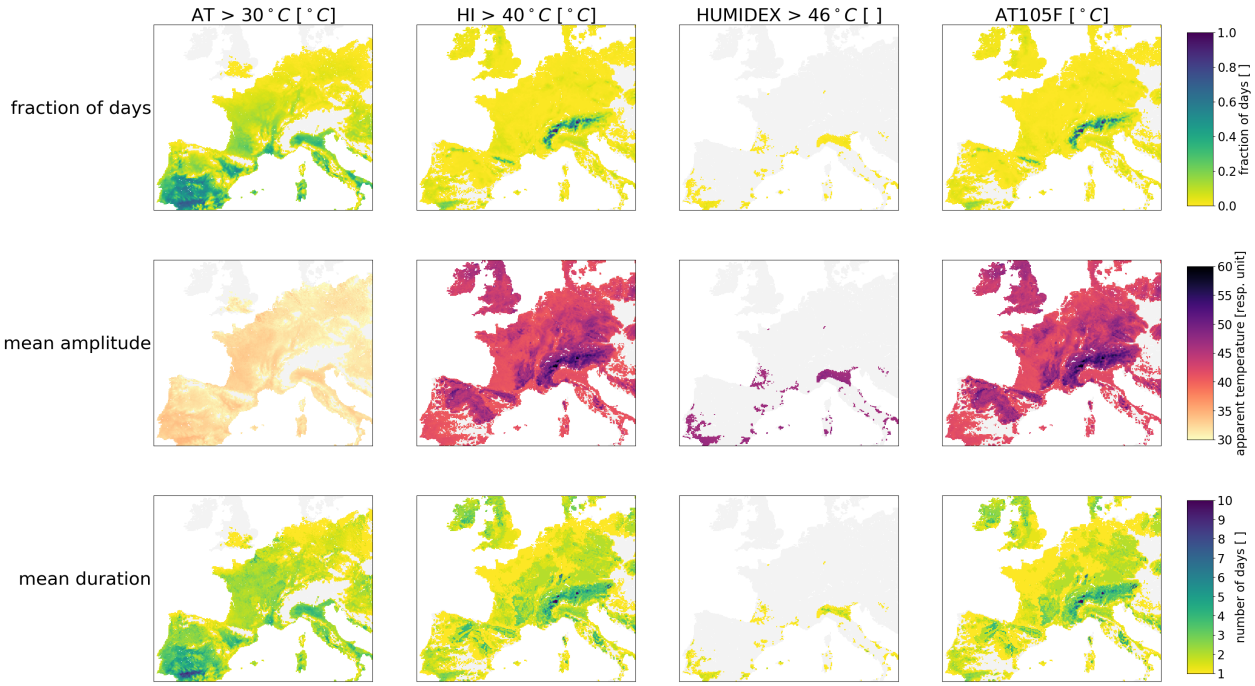
4 Discussion

The results show differing model behaviour at the land surface in the three summers. In the following, the important points of both findings are expanded. The differences in surface energy partition in the two models are characterised and linked to observed temperature biases and LSM structure. The differences between the results for 2-meter temperature and radiative temperature are discussed. Furthermore, the representation of human comfort indices in both models is compared and evaluated. And finally, limitations and uncertainties in this study are discussed.

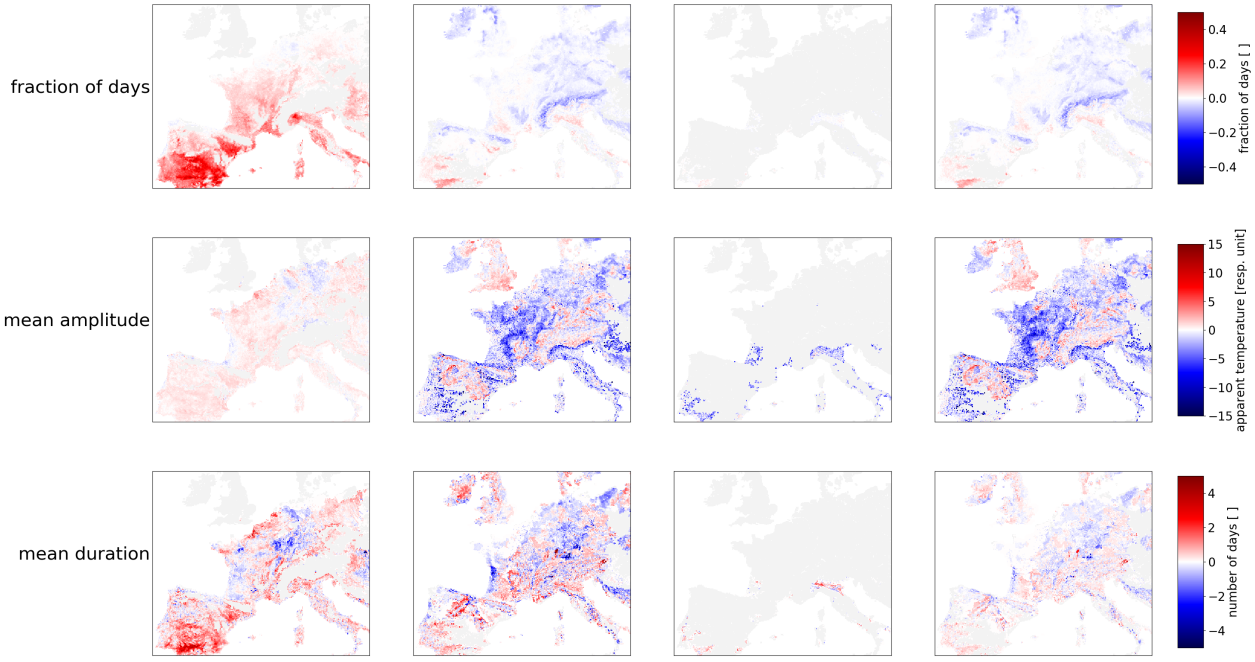
4.1 Surface Energy Balance

Both models overestimate sensible heat flux and analogously underestimate latent heat flux during the summer months. This behaviour explains the warm bias in both models when compared to EOBS: The ratio of sensible to latent heat is distorted towards too much sensible heat and too little latent heat, warming the air at 2-meter level. While COSMO-CLM has the same sign of bias as COSMO-TERRA, the magnitude is smaller and more similar between months.

The misrepresentation of surface fluxes is three times higher in July in Western Europe for COSMO-TERRA than in other months. This coincides in time and space with the first major heat event



(a) COSMO-TERRA



(b) COSMO-TERRA - COSMO-CLM

Figure 17: Magnitude of heat extremes exceeding human comfort thresholds over Europe for all three summers expressed as fraction of days affected (upper panels), mean amplitude of heat extremes (middle panels) and mean duration of heat extremes (lower panels) for (a) COSMO-TERRA and (b) the difference between COSMO-TERRA and COSMO-CLM (COSMO-TERRA - COSMO-CLM). For COSMO-CLM see Supplementary Figure A12. Indices used here are, from left to right, apparent temperature after Buzan et al. (2015), Heat Index, HUMIDEX and AT105F (see Fischer and Schär, 2010).

hitting Western Europe end of June to beginning of July (see Figure 1c) and initiates the departure of COSMO-TERRA evaporative fraction from observations and COSMO-CLM towards values around 40%. In hot summers COSMO-TERRA overestimates sensible heat and, subsequently, temperature. In contrast, the overestimation of sensible heat flux in COSMO-CLM is constant in time and COSMO-CLM is able to keep evaporative fraction at observed values. Davin et al. (2016) found a similar drop in latent heat starting in June for COSMO-TERRA, suggesting a limited control of vegetation on evapotranspiration in the beginning of summer and a subsequent anomalous dry soil in summer. They used COSMO version 4.8, but Supplementary Figure A17 shows COSMO version 5.0 has comparable results for both LSMs.

However, the soil moisture anomaly (as compared to values on the beginning of the year) in the root zone in both LSMs is similar and can not explain the drop in evaporative fraction in COSMO-TERRA. Nevertheless, COSMO-TERRA is on absolute values drier than COSMO-CLM, possibly because of soil texture effects or the missing groundwater treatment. Soil moisture in COSMO-TERRA might drop below field capacity for larger regions where it can not be used for transpiration by the plants anymore or the empirical transpiration parameterisation in COSMO-TERRA is biased towards underestimating transpiration for low soil moisture.

Its not intuitive why the superior flux representation in COSMO-CLM did not translate into similar temperature scores. Since both models share the same tuning which is optimized for COSMO-TERRA, this calls for a retuning of COSMO-CLM to redo the analysis and shows the importance of model tuning. Improved temperature scores in COSMO-CLM after retuning would confirm this hypothesis.

The benchmark latent heat estimate has better scores than both models, a considerable difference when compared to COSMO-TERRA and a small difference when compared to COSMO-CLM. A simple statistical model hence is able to outperform both models even in these extremely hot summers. This suggests information usage of both models can be considerably enhanced. The lower variability of the benchmark compared to GLEAM is attributed to the linear assumption of the underlying statistical model, that cannot fully modulate observed variability.

Davin et al. (2016) found interannual summer temperature variability to be smaller in COSMO-CLM and closer to observations and attributed this to the enhanced seasonal cycle of ground heat flux in COSMO-CLM, allowing for more energy to be stored in the ground in summer and released in winter, dampening the seasonal cycle. They attribute this to the much deeper bottom boundary condition for thermal processes in CLM. In agreement with these results, we find that the daily averaged ground heat flux in COSMO-CLM exhibits a larger seasonal cycle.

However, the magnitude of the ground heat flux estimated from evaluation datasets is 20 - 40 Wm^{-2} larger than the modelled flux, while ground heat flux at Lindenberg is smaller than mod-

elled flux. It would be advantageous to be able to directly obtain the ground heat flux as estimated by the model and not calculate it as a residual form all the other fluxes. There is no indication to assume that the datasets used for the ground heat flux estimate have accumulating biases, so this difference remains unexplained. The disagreeing variance, however, may stem from adding the variance of different datasets. Erratic peaks in ground heat flux during winter in models and evaluation estimate could result from snow melting effects.

When looking at sub-daily time scales, the shape of the diurnal cycle of the ground heat flux differs substantially among models. Ground heat flux is estimated as a residual of all other fluxes: Sensible heat flux in COSMO-TERRA is on average over 100 Wm^{-2} larger than in COSMO-CLM, especially in summer and the southern parts of Europe, while in COSMO-CLM more heat is stored in the ground. This larger sensible heat flux in the morning leads to a smaller latent heat flux in the afternoon in COSMO-TERRA, when stores more heat in the ground. COSMO-CLM, in contrast, releases much of is heat stored in the ground in the afternoon. Heat flux into the ground is thus larger in COSMO-CLM in the morning and heat flux back into the atmosphere is larger in the afternoon compared to COSMO-TERRA. Since the diurnal cycle of both models has more similarity between years than between models, we argue this is a robust result.

The release of ground heat flux in COSMO-CLM in the afternoon is one possible explanation why the temperature peak at 2 meters in COSMO-CLM is on average one hour later than in COSMO-TERRA. Another reason could be that in COSMO and CLM are only coupled every 20mins, so CLM experiences the temperature change in COSMO with a 20mins delay and adjusts the fluxes at the surface accordingly, leading to a delayed diurnal temperature development.

The ground heat flux at Lindenberg is considerably *smaller* than the modelled ground heat flux. However, the comparison is hampered since ground heat flux at Lindenberg is measured at 5 cm depth, where we already expect considerable dampening of the signal. Furthermore, the vegetation in the models at Lindenberg is different from the vegetation at the meteorological station, restricting comparability of model and station data. However, when correcting for the measurement depth and the vegetation at Lindenberg, Schulz and Vogel (2017) also found the modelled ground heat flux to be larger.

Comparing both models, Schulz and Vogel (2017) found an increased diurnal cycle in ground heat flux in COSMO-TERRA at Lindenberg, while in COSMO-CLM the flux is smaller and closer to observations. They argue missing shading from vegetation and missing vegetation heat budget increase heat flux into the ground in COSMO-TERRA. In CLM, with shading and vegetation temperature introduced, this bias is smaller. With the new formulation of radiative "skin" temperature in COSMO-TERRA version 5.05 (see Section 4.5), these biases are expected to be reduced. However, in contrast to Schulz and Vogel (2017) we find an enlarged diurnal cycle in ground heat flux

in COSMO-CLM as compared to COSMO-TERRA at Lindenberg.

Sensible and latent heat flux at Lindenberg show similar behaviour in both models than over the whole domain. However, Schulz and Vogel, 2017 found the evaporative fraction at Lindenberg too large, while our comparison with gridded datasets and Lindenberg data suggest underestimation of evaporative fraction.

The error in the latent heat estimation in COSMO-TERRA is scaling with the LAI, i.e. the more vegetation in a grid cell prevails, the less accurate is the latent heat estimation in COSMO-TERRA. COSMO-CLM does not show a dependency of latent heat misrepresentation with vegetation properties. Two candidate processes for this error in COSMO-TERRA exist: The parameterization for plant transpiration or the bare soil evaporation. Since available COSMO-TERRA output does not distinguish between sources of latent heat flux, this question cannot be answered in this thesis. However, it is an interesting pathway to investigate. Both GLEAM and COSMO-CLM distinguish between transpiration and evaporation, the availibility of comparable observations and model output is therefore warranted.

The difference between latent heat estimates of GLEAM and WECANN stems from their different spatial and especially temporal time resolution. WECANN only provides monthly estimates, triggering better performance in models and benchmark as compared to GLEAM, since monthly averages are easier to get right than daily averages.

Unfortunately, GLEAM, WECANN and the Lindenberg station do not provide error estimates, so an assessment of measurement uncertainty in comparison to reported bias is not possible. The model runs, since not run as ensembles, do also not provide error estimates. However, when comparing the different summers, the results always agree more between summers than between models, which gives confidence for the reported model bias.

4.2 Temperature Bias

Warm bias of 2-meter mean temperature in COSMO-TERRA and COSMO-CLM has already been shown for climate runs in different model setups (Davin et al., 2016, Davin and Seneviratne, 2012, Lorenz et al., 2012). Davin et al. (2016) also found the bias to be larger in COSMO-TERRA than in COSMO-CLM. Daily average temperatures in both models are represented best, which is not surprising since this is the temperature the models are tuned for.

Analogue to Davin et al. (2016), we find that COSMO-CLM outperforms COSMO-TERRA in 2-meter mean and minimum temperature. However, for 2-meter maximum temperature, where Davin et al. (2016) found the largest improvement in COSMO-CLM, we find COSMO-TERRA outper-

forms COSMO-CLM because COSMO-CLM is too cold. Both models show a small cold bias during the day. Observed maximum temperatures had higher anomalies than minimum temperatures in July 2006 and summer 2003 (Rebetez et al., 2009), possibly challenging the capability of both models to correctly simulate daily maximum temperatures.

Radiative temperature overestimation in Western France and the UK and underestimation in the Mediterranean and in the Alps is consistent among summers (see also Supplementary Figures A4 and A5). Both models show considerable 3-4 K cold bias during day in the Southern European domain. COSMO-CLM additionally is on average around 1 K colder than COSMO-TERRA. Mean ground temperature and, more severly, mean vegetation temperature in COSMO-CLM have widespread cold biases and cannot outperform the ground temperature from COSMO-TERRA.

Underestimation of daily maximum radiative temperature is most pronounced on the Iberian Peninsula. We attribute this to the outdated aerosol climatology of Tanre et al. (1984) in COSMO, overestimating Saharan dust influx into the Mediterranean and subsequently underestimating shortwave radiation by more than 35 Wm^{-2} (Zubler et al., 2011). The difference in aerosol optical depth between the climatology of Tanre et al. (1984) and observations is especially large on the Iberian Peninsula (see Zubler et al., 2011, Figure 1b, f). An underestimation of shortwave radiation in both models has already been found by Davin et al. (2016) and a smaller underestimation is visible here for the Mediterranean area (see Supplementary Figures A8, A6 and A7). A perfect LSM would therefore show a consistent cold bias, most severe on the Iberian Peninsula. Updating this climatology in the COSMO version used for weather forecast should therefore be a priority in COSMO development to examine whether the constant cold bias in COSMO-CLM is partly due to this aerosol treatment.

The mean diurnal cycle of temperatures in both models is reduced in similar magnitude. Daily maximum temperatures at 2 meters are too small in both models over the whole European domain, with COSMO-CLM having colder temperatures than COSMO-TERRA.

For radiative temperatures, the reduced diurnal cycle stems from a large (on average 5 K) cold bias during the day in southern Europe. Pronounced cold biases in modelled surface temperatures, when compared to satellite-retrieved land surface temperature estimates, are known (see Zheng et al., 2012, Garand, 2003) and attributed to misrepresentation of fluxes at the surface, mainly stemming from simplistic model parameterisation of radiative temperature and little knowledge of land cover and soil properties (Trigo et al., 2015). Trigo et al. (2015) also found this feature to be most prominent in semi-arid regions, similar to our analysis where we find cold bias mostly in Southern Europe (see Figure 13c, d).

Schulz and Vogel (2017) argue an updated, increased soil resistance in the bare soil evaporation introduced in COSMO version 5.05 could substantially reduce this bias by reducing latent heat

flux during the day and therefore increasing daily maximum temperatures. Additionally, a more realistic leaf phenology increases LAI in spring and autumn and so would also enhance latent heat flux. They find these improvement enhance daily maximum temperatures when compared to the Lindenberg station. However, since we find the latent heat flux to be underestimated in COSMO-TERRA and smaller than in COSMO-CLM, we do not expect this new bare soil evaporation scheme to improve the results. Our results counteract the ones found in Schulz and Vogel (2017), calling for a more in-depth analysis.

Both models are too warm at 2 meters at night over the whole European domain. This feature is known for COSMO and also common among other RCMs (Davin et al., 2016). Misrepresentation of nighttime stratification is a candidate process for this nocturnal warm bias.

Nighttime bias in radiative temperatures differs among models, COSMO-CLM is too cold and COSMO-TERRA too warm. Schulz and Vogel (2017) also found a reduced nocturnal warm bias of radiative temperature in COSMO-CLM compared to COSMO-TERRA and attributed it to the smaller ground heat flux they found in COSMO-CLM at Lindenberg. However, since ground heat flux at Lindenberg and throughout the domain is larger in COSMO-CLM than in COSMO-TERRA in this study, we cannot argue similarly. A new ”skin” conductivity formulation introduced in COSMO-TERRA version 5.05 has shown to significantly reduce ground heat flux and subsequently nighttime warm bias at Lindenberg (Schulz and Vogel, 2017). Smaller ground heat flux into the ground during day and warming the air during night was similarly achieved by reducing conductivity of the vegetation ”skin” in a different LSM in Trigo et al. (2015). Integrating the new COSMO-TERRA version into the ground heat flux analysis of this study is crucial to shed more light on the discrepancies between the results of this study and the experiments in Schulz and Vogel (2017).

Although daily maximum temperature at 2-meters is better represented in COSMO-TERRA, the representation of maximum-temperature-only heat extreme indices is similar in both models. This suggests a better performance of COSMO-CLM with extreme maximum daily temperatures. With indices also considering nighttime temperatures, COSMO-CLM is outperforming COSMO-TERRA since it overestimates nighttime temperatures less. Night-time temperatures have shown to also be an important variable to weather-related deaths (Poumadère et al., 2005). Therefore COSMO-CLM has an advantage here, since it outperforms COSMO-TERRA when nighttime temperatures are taken into account.

Potential SLST retrieval errors are included in the SLST dataset. In Figure 12 pixels where the retrieval error of the SLST estimate is larger than the reported bias are stippled. This happens only in areas where the bias is close to zero, giving confidence to the assessment of the reported areas of large biases.

4.3 Limitations of Radiative Temperature Evaluation

Radiative temperature and 2-meter temperature are two distinctively different temperature variables. Radiative temperature represents the temperature of the land surface heated up by incoming radiation and is directly estimated by surface energy partition. Radiative temperature therefore exhibits a larger diurnal cycle, easily reaching up to 50 K during a hot day and cooling up fast once insolation has stopped. Additionally, the SLST dataset only samples cloud-free conditions, where higher temperature amplitudes are expected. 2-Meter temperature, on the other hand, shows a more smooth diurnal cycle.

The inconsistencies among biases of both temerature variables shown in Section 3.2.3 can stem from the definition of both temperatures in the model or inconsistencies in the observational datasets. Results from the comparison with EOBS agree well with already published literature (see for example Davin et al., 2016), but the interpolation of irregularly distributed and partially sparse observational data does not come without limitation. Uncertainty in interpolated values increases with distance to the next measurement stations and the complexity of the terrain (Haylock et al., 2008).

Model verification with satellite-derived land surface temperature measurements has shown to be challenging (see Zheng et al., 2012, Garand, 2003, Wang et al., 2014) because of simplistic radiative temperature definitions in models and disagreement in land cover between model and observations. In COSMO-TERRA, SLST comparison is hampered by the fact that T_{rad} is defined *below* the vegetation, i.e. without influence of the vegetation. In CLM, two temperatures are available for comparison (T_{rad} and T_v, see Section 2.2). In contrast, Trigo et al. (2015) use a LSM that defines "skin" temperature of the LSM above vegetation for vegetated areas and directly on the bare soil for unvegetated areas. Such a definition of LSM "skin" temperature is more readily comparable to satellite-retrieved land surface temperature estimates.

Satellite-derived land surface temperatures have also shown to have a larger error with higher temperatures (see Duguay-Tetzlaff et al., 2015, Figure 4b), a feature common also for similar datasets (Heidinger et al., 2013). Duguay-Tetzlaff et al. (2015) estimate the error in summer to be 3-4 K, comparable to the magnitude of the bias reported in this study.

The SLST dataset only samples cloud-free conditions. Cloud cover in model is not necessarily the same as in SLST, but not available for analysis. In this analysis, only points clouded in the satellite data are removed from analysis. Where cloudy conditions in the model prevail, but the satellite sees cloud-free sky, higher temperature amplitudes are measured by the satellite. However, we performed a sensitivity test by repeating the analysis with model data only considered where the cloud cover fraction is below 20% for the year 2003 – the only year where the cloud cover variable is available in model output – and see little difference (see Figure A19). We therefore argue the

cloud cover in model and SLST is similar and has negligible effect on our results.

Additionally, the satellite data gives instantaneous values at each full hour, while model data reports the average of the previous hour at each full hour. With fast-changing conditions occuring in the hour before, these two variables do not represent the same conditions. Furthermore, the uncertainty of the satellite measurement increases towards the edge of the observational disk visible by the satellite that is located in a geostationary orbit on the equator. Since we do not see an increase in error towards Northern Europe, we argue this effect is negligible.

4.4 Heat Extreme Representation

Limitations and inaccuracies of human comfort indices are discussed (see for example Buzan et al., 2015), but especially heat extremes combined with high humidity are known to have impact on human health (McGregor et al., 2015). Therefore it is inherently important to look beyond temperature to measure heat stress for humans. Even if, for example in simple indices with COSMO-TERRA, extremeness is overestimated, while underestimation would be more dangerous, it is still important to get these indices correct since it is crucial for models to not be right for the wrong reasons.

We cannot evaluate model performance for human comfort indices, since observations are lacking. Variability is high among human comfort indices and agreement low, which we attribute to their substantially different definitions hampering comparison between indices. HI40 and AT105F show higher apparent temperatures in France for COSMO-CLM. In contrast, AT30 events occur more often, are more prolonged and slightly warmer in COSMO-TERRA especially on the Iberian Peninsula and in France. AT30 has overall lower amplitudes, since wind-speed is taken into account explicitly and considerably lowers apparent temperature estimates. A troublesome feature is that indices disagree most in areas with high reported apparent temperatures.

The detection of large-scale heat extremes for HI40 and AT105F in the Alps stems from the reported humidity, which in the summers in on average is above 80% around the Alpine region in both models (see Supplementary Figure A18). We argue HI and AT105F both are linearly regressed estimates of apparent temperature and include relative humidity in a quadratic fashion. When relative humidity is large and temperatures are low, as in the Alpine region, these indices weigh relative humidity at the expense of temperature, leading to the detection of heat extremes under low temperatures. Such indices are therefore not suitable for domains with considerable orograhy.

4.5 Limitations and Uncertainties

Models often disagree with observations in the Alpine region and in coastal areas. For the coastal areas, we attribute this to small inconsistencies in the land mask of models and observations. Since observed variables change sharply at the land-sea border, errors are large when one datasets assumes the area to be land and the other ocean. This is visible for example in Figure 11.

In regions of high topography as in the Alps, both evaluation datasets and models need altitude correction. The lower the resolution in the model, the lower are mountains in the model. EOBS depends on station data and the representativeness of the station on the surrounding area decreases with increasing complexity of the terrain. The EOBS interpolation algorithm more randomly smoothes between stations in the Alps compared to less orographic terrain (Hofstra et al., 2010). SLST corrects for atmospheric conditions using ERA-Interim height, whose low resolution also underestimates Alpine topography. Furthermore, satellite observed temperature sees shading of mountains in areas of high topography. These uncertainties are not represented in the error estimate of SLST. Only if both evaluation dataset and model assume the same height at a point in areas of large topography, the estimates can be correctly compared. We therefore exclude results in the Alpine region from our analysis assuming measurement and comparison errors to be considerable.

Apart from challenges in mountaneous regions, each evaluation dataset comes with its own uncertainties. Except from SLST, none of the other evaluation datasets has an error estimate included in this analysis. However, seeing similar results in the three years gives us confidence the reported biases are robust for hot summers.

In the proposal of this master thesis, several objectives were outlined that did not find their way into this thesis. This has several reasons. From twelve planned model runs, only six were finished in time to be incorporated in this master thesis. Originally, COSMO-TERRA was planned to be run not only in the currently-used version 5.0, but also in the newly developed version 5.05, both with the the tuning of 5.0 and a new tuning for 5.05 for all three summers.

However, COSMO-TERRA version 5.05 release was delayed to February this year (see release note), too late to perform the model runs to be included here. COSMO-TERRA 5.05 includes new developments on bare soil evaporation and "skin" temperature formulation which are important in the context of hot summers. A follow-up project is therefore necessary to assess the improvements of these newly implemented features for land surface representation in COSMO. The overarching TERRA-NOVA project is elongated until the end of 2018 to allow for the project to finish the standard verification of the new COSMO-TERRA version 5.05.

Furthermore, the standard verification from MeteoSwiss is missing due to delays from the TERRA-

NOVA project. The verification of both models with standard statistical measures (see Table 3) therefore had to be performed from scratch with the available datasets.

Additionally, some variables were missing in the output. Radiative temperature comparison could not take into account cloud cover in the model, since the cloud cover variable was not available in the model runs. One available run with cloud cover could be used for sensitivity testing, showing that the difference is small (see Supplementary Figure A19). Incoming shortwave radiation was not available for the benchmark experiment, which is why we used net shortwave radiation instead. T_v in CLM was only available on daily, not hourly timesteps. The analysis steps that were hampered by missing output variables from the first batch of runs could be redone in a possible follow-up project to ascertain the uncertainties involved with these findings.

Because of this additional workload and the delayed runs of the two models compared in this thesis, the last objective on sensitivity analysis of parameterisations was dropped. A follow-up project could perform these tasks and also incorporate runs with the new COSMO-TERRA version 5.05 as well as the standard verification of MeteoSwiss.

5 Conclusions

Overshooting sensible heat production at the expense of evapotranspiration in COSMO-TERRA is consistent among all hot summers. COSMO-CLM has biases of similar sign, but smaller magnitude for land surface fluxes during summer heat extremes. Additionally, COSMO-TERRA increases in error of latent heat estimation with LAI and summer heat, while COSMO-CLM is more robust. This reduces trust in COSMO-TERRA during hot summers, pinpointing towards structural LSM deficiencies in the representation of transpiration by plants or bare soil evaporation in the LSM TERRA. Since observational datasets on evaporation and transpiration are available (see Martens et al., 2017) and can be distinguished in CLM, it is possible to further investigate which of these processes in TERRA is responsible.

The statistical benchmark is able to outperform both models in latent heat estimation with regard to error and correlation. The difference is most pronounced for COSMO-TERRA. Such benchmark experiments give new valuable insights of information usage of the LSMs and disclose leeway in model improvements.

Enhanced surface energy balance representation in COSMO-CLM did, quite non-intuitively, not translate into a better temperature representation. COSMO-CLM is constantly colder than COSMO-TERRA, most pronounced in the Mediterranean. This cold bias is beneficial to model performance compared with EOBS, but degrades model performance compared to SLST. COSMO-CLM out-

performs COSMO-TERRA at 2 meters except for maximum temperatures, and COSMO-TERRA outperforms COSMO-CLM for radiative temperatures, except for minimum temperatures. The models share a model tuning tailored to COSMO-TERRA, an advantage over COSMO-CLM. Rerunning both models with an individual tuning for each model could potentially lift deficiencies of COSMO-CLM and reduce discrepancies to the results found in Davin et al. (2016) in climate mode.

Additionally, a substantial cold bias by the outdated aerosol treatment in COSMO of Tanre et al. (1984) is induced, overestimating Saharan dust influx into the Mediterranean and subsequently severly underestimating incoming shortwave radiation (Zubler et al., 2011). With this aerosol treatment in place, a perfect LSM would still show a persistent cold bias, most pronounced in the summer. Implementing a more realistic aerosol treatment (e.g. Tegen et al., 1997) could enhance performance especially of COSMO-CLM. A possible follow-up project with a new aerosol treatment, individual model tunings and additional output variables could resolve remaining questions.

Summer heat extremes can be dangerous to human health. Correctly representing their impact on humans does not only include taking daily maximum temperatures into account. The impact on human health also depends on daily minimum temperatures (Poumadère et al., 2005) and humidity (McGregor et al., 2015). COSMO-CLM outperforms COSMO-TERRA for simple indices of temperature extremes although temperature representation is hampered. Human comfort indices show mostly larger values for COSMO-CLM, but have troublesome disagreement especially in areas where heat is extreme. Comparison to observations is necessary here, but was not possible in this work because of missing relative humidity estimates.

The seasonal cycle of ground heat flux in both models is underestimated compared to gridded observations but overestimated on seasonal *and* diurnal time scales compared to the meteorological station at Lindenberg, Germany. The seasonal cycle of sensible heat flux is overestimated and the seasonal cycle of latent heat is underestimated compared to Lindenberg (in agreement with findings over the whole domain, but in disagreement with Schulz and Vogel, 2017). Ground heat flux in COSMO-CLM is larger than COSMO-TERRA on seasonal cycles (in agreement with Davin et al., 2016) and on diurnal cycles both on domain average and at Lindenberg station (in disagreement with Schulz and Vogel, 2017). Disagreeing results when comparing to station data and gridded data are troublesome and call for a further investigation, especially since LSM development largely relies on verification with station data.

Replacing COSMO-TERRA with COSMO-CLM for weather forecast at DWD and MeteoSwiss is not planned, since control on LSM development would be stripped (CLM is maintained at NCAR) and the highly resolved surface input fields required for CLM are difficult to obtain at the high 1 km resolution of their forecast model.

However, planned improvements in parameterisation for COSMO-TERRA version 5.05 can be evaluated using the results of this study. A new ”skin” temperature formulation, which should reduce diurnal cycle in ground heat flux to reduce nocturnal warm bias in COSMO-TERRA is promising. We argue ”skin” temperature formulation is most important, along with changing the aerosol treatment. A new resistance-based bare soil evaporation in COSMO-TERRA version 5.05 should reduce latent heat. However, we see underestimation of latent heat in COSMO-TERRA on domain average. This questions how this development would impact on overall model performance.

Incorporating satellite-derived observations of land surface temperature into model verification and data assimilation is hampered by simplistic radiative temperature definitions in models and disagreement on land cover between satellite and model (Trigo et al., 2015). A pronounced daytime cold bias and a smaller nighttime warm bias in models is common when compared to satellite-derived products (see Zheng et al., 2012, Garand, 2003) and attributed to a misrepresented surface energy balance (Trigo et al., 2015).

We argue a biased surface energy balance most pronounced in COSMO-TERRA hampers comparison with the SLST dataset, alongside considerable 3-4 K uncertainties in the observations that are just little smaller than the observed bias in the models. Additionally, differing definitions of radiative temperature between COSMO-TERRA, COSMO-CLM and SLST limit comparability. Especially in COSMO-TERRA, radiative temperature is independent of the vegetation layer, a major simplification which is planned to be partially lifted by the new ”skin” temperature formulation and a new phenology of COSMO-TERRA version 5.05.

Improvement in weather forecast by incorporating LST products has been shown for other models (Orth et al., 2017). Their high resolution and location directly at the land surface energy balance can potentially have a huge impact on LSM performance, since surface energy balance is a crucial feature of LSM performance and a measure of the land surface coupling that is especially important during hot summers. Thus, including LST products into future model development and verification is a crucial next step. Orth et al. (2017) even argue the utilization of LST products to overcome model shortcomings could potentially help LSMs to finally outperform well calibrated statistical models, a limitation of LSMs also shown in this study. However, more physical parameterisations do not necessarily lead to model improvement. This intrinsic trade-off between model development and model performance poses a challenge to model development, since a model can always be right for the wrong reasons.

References

Abramowitz, G., Leuning, R., Clark, M., and Pitman, A. (2008). Evaluating the Performance of Land Surface Models. *Journal of Climate*, 21(21):5468–5481.

Alemohammad, S. H., Fang, B., Konings, A. G., Aires, F., Green, J. K., Kolassa, J., Miralles, D., Prigent, C., and Gentine, P. (2017). Water, Energy, and Carbon with Artificial Neural Networks (WECANN): a statistically based estimate of global surface turbulent fluxes and gross primary productivity using solar-induced fluorescence. *Biogeosciences*, 14(18):4101–4124.

Beniston, M. (2004). The 2003 heat wave in Europe: A shape of things to come? An analysis based on Swiss climatological data and model simulations. *Geophysical Research Letters*, 31(2).

Beniston, M., Stephenson, D. B., Christensen, O. B., Ferro, C. A. T., Frei, C., Goyette, S., Halsnaes, K., Holt, T., Jylhä, K., Koffi, B., Palutikof, J., Schöll, R., Semmler, T., and Woth, K. (2007). Future extreme events in European climate: an exploration of regional climate model projections. *Climatic Change*, 81(1):71–95.

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., van den Hurk, B. J. J., Nearing, G. S., Pak, B., Peters-Lidard, C., Santanello, J. A., Stevens, L., and Vuichard, N. (2015). The Plumbing of Land Surface Models: Benchmarking Model Performance. *Journal of Hydrometeorology*, 16(3):1425–1442.

Bettems, J.-M. (2017). Work Group 3b: Physical aspects, Soil and Surface. http://www.cosmo-model.org/content/tasks/workGroups/wg3b/default.htm. Last checked on May 30, 2015.

Beven, K. J., Kirkby, M. J., Schofield, N., and Tagg, A. F. (1984). Testing a physically-based flood forecasting model (TOPMODEL) for three U.K. catchments. *Journal of Hydrology*, 69(1):119–143.

Buzan, J. R., Oleson, K., and Huber, M. (2015). Implementation and comparison of a suite of heat stress metrics within the Community Land Model version 4.5. *Geosci. Model Dev.*, 8(2):151–170.

Collatz, G., Ribas-Carbo, M., and Berry, J. (1992). Coupled Photosynthesis-Stomatal Conductance Model for Leaves of C 4 Plants. *Functional Plant Biology*, 19.

Collatz, G. J., Ball, J. T., Grivet, C., and Berry, J. A. (1991). Physiological and environmental regulation of stomatal conductance, photosynthesis and transpiration: a model that includes a laminar boundary layer. *Agricultural and Forest Meteorology*, 54(2):107–136.

COPA COGECA (2003). Assessment of the impact of the heat wave and drought of the summer 2003 on agriculture and forestry. Technical report, Committee of Agricultural Organisations in the European Union General Committee for Agricultural Cooperation in the European Union, Brussels. http://docs.gip-ecofor.org/libre/COPA_COGECA_2004.pdf. Last checked on May 30, 2015.

Davin, E. L., Maisonnave, E., and Seneviratne, S. I. (2016). Is land surface processes representation a possible weak link in current Regional Climate Models? *Environmental Research Letters*, 11(7):074027.

Davin, E. L. and Seneviratne, S. I. (2012). Role of land surface processes and diffuse/direct radiation partitioning in simulating the European climate. *Biogeosciences*, 9(5):1695–1707.

Davin, E. L., Stöckli, R., and Jaeger, E. B. (2011). COSMO-CLM2: a new version of the COSMO-CLM model coupled to the Community Land Model. *Climate Dynamics*, 37(9-10):1889–1907.

De Bono, A., Peduzzi, P., Kluser, S., and Giuliani, G. (2004). Impacts of summer 2003 heat wave in Europe. *UNEP Environment Alert Bulletin*.

Dickinson, R. E. (1984). Modeling evapotranspiration for three-dimensional global climate models. *Washington DC American Geophysical Union Geophysical Monograph Series*, 29:58–72.

Doms, G., Forstner, J., Heise, E., Reinhardt, T., Ritter, B., and Schrodin, R. (2011). A Description of the Nonhydrostatic Regional COSMO Model. *Deutscher Wetterdienst*, page 161.

Donat, M. G., Pitman, A. J., and Seneviratne, S. I. (2017). Regional warming of hot extremes accelerated by surface energy fluxes: Accelerated Warming of Hot Extremes. *Geophysical Research Letters*, 44(13):7011–7019.

Dougherty, R. L., Bradford, J., Coyne, P. I., and Sims, P. L. (1994). Applying an empirical model of stomatal conductance to three C-4 grasses. *Agricultural and Forest Meteorology*, 67.

Duguay-Tetzlaff, A., Bento, V., Göttsche, F., Stöckli, R., Martins, J., Trigo, I., Olesen, F., Bojanowski, J., da Camara, C., and Kunz, H. (2015). Meteosat Land Surface Temperature Climate Data Record: Achievable Accuracy and Potential Uncertainties. *Remote Sensing*, 7(12):13139–13156.

Farquhar, G. D., Caemmerer, S. v., and Berry, J. A. (1980). A biochemical model of photosynthetic CO2 assimilation in leaves of C3 species. *Planta*, 149(1):78–90.

Ferranti, L. and Viterbo, P. (2006). The European Summer of 2003: Sensitivity to Soil Water Initial Conditions. *Journal of Climate*, 19(15):3659–3680.

Fink, A. H., Brücher, T., Krüger, A., Leckebusch, G. C., Pinto, J. G., and Ulbrich, U. (2006). The 2003 European summer heatwaves and drought –synoptic diagnosis and impacts. *Weather*, 59(8):209–216.

Fischer, E. M. and Schär, C. (2010). Consistent geographical patterns of changes in high-impact European heatwaves. *Nature Geoscience*, 3(6):398–403.

Fischer, E. M., Seneviratne, S. I., Lüthi, D., and Schär, C. (2007a). Contribution of land-atmosphere coupling to recent European summer heat waves. *Geophysical Research Letters*, 34(6).

Fischer, E. M., Seneviratne, S. I., Vidale, P. L., and others (2007b). Soil Moisture–Atmosphere Interactions during the 2003 European Summer Heat Wave. *Journal of Climate*, 20(20):5081–5099.

Garand, L. (2003). Toward an Integrated Land–Ocean Surface Skin Temperature Analysis from the Variational Assimilation of Infrared Radiances. *Journal of Applied Meteorology*, 42(5):570–583.

García-Herrera, R., Díaz, J., Trigo, R. M., Luterbacher, J., and Fischer, E. M. (2010). A Review of the European Summer Heat Wave of 2003. *Critical Reviews in Environmental Science and Technology*, 40(4):267–306.

Grasselt, R., Schuettemeyer, D., Warrach-Sagi, K., Ament, F., and Simmer, C. (2008). Validation of TERRA-ML with discharge measurements. *Meteorologische Zeitschrift*, 17:763–773.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., and New, M. (2008). A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *Journal of Geophysical Research: Atmospheres*, 113(D20):D20119.

Heidinger, A. K., Laszlo, I., Molling, C. C., and Tarpley, D. (2013). Using SURFRAD to Verify the NOAA Single-Channel Land Surface Temperature Algorithm. *Journal of Atmospheric and Oceanic Technology*, 30(12):2868–2884.

Held, I. M. (2005). The Gap between Simulation and Understanding in Climate Modeling. *Bulletin of the American Meteorological Society*, 86(11):1609–1614.

Hofstra, N., New, M., and McSweeney, C. (2010). The influence of interpolation and station network density on the distributions and trends of climate variables in gridded daily data. *Climate Dynamics*, 35(5):841–858.

Jaeger, E. B. and Seneviratne, S. I. (2011). Impact of soil moisture–atmosphere coupling on European climate extremes and trends in a regional climate model. *Climate Dynamics*, 36(9-10):1919–1939.

Knutti, R., Masson, D., and Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there: CLIMATE MODEL GENEALOGY. *Geophysical Research Letters*, 40(6):1194–1199.

Koffi, B. and Koffi, E. (2008). Heat waves across Europe by the end of the 21st century: Multiregional climate simulations. *Climate Research*, 36.

Kovats, R. and Hajat, S. (2008). Heat stress and public health: A critical review. *Annual Review of Public Health*, 29:41–55.

Lawrence, P. J. and Chase, T. N. (2007). Representing a new MODIS consistent land surface in the Community Land Model (CLM 3.0). *Journal of Geophysical Research*, 112(G1).

Lorenz, R., L. Davin, E., and Seneviratne, S. (2012). Modeling land-climate coupling in Europe: Impact of land surface representation on climate variability and extremes. *Journal of Geophysical Research (Atmospheres)*, 117:20109.

Luterbacher, J., Dietrich, D., Xoplaki, E., Grosjean, M., and Wanner, H. (2004). European Seasonal and Annual Temperature Variability, Trends, and Extremes Since 1500. *Science*, 303(5663):1499–1503.

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *MULTIVARIATE OBSERVATIONS*, page 17.

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E. C. (2017). GLEAM v3: satellite-based land evaporation and root-zone soil moisture. *Geosci. Model Dev.*, 10(5):1903–1925.

Masterton, J. M. (1979). *Humidex: A Method of Quantifying Human Discomfort Due to Excessive Heat and Humidity, by J.M. Masterton and F.A. Richardson.* Ministere de l'Environnement. Google-Books-ID: qzPbOAAACAAJ.

McGregor, G. R., Bessemoulin, P., Ebi, K., and Menne, B. (2015). WHO | Heatwaves and health: guidance on warning-system development. http://www.who.int/globalchange/publications/heatwaves-health-guidance/en/. Last checked on May 30, 2015.

NOAA (2018). Climate at a glance: Global time series. https://www.ncdc.noaa.gov/cag/global/time-series. Last checked on May 31, 2015.

Oleson, K. W. and Lawrence, D. M. (2013). Technical Description of version 4.5 of the Community Land Model (CLM). *NCAR Technical Note*, page 434.

Orth, R., Dutra, E., Trigo, I. F., and Balsamo, G. (2017). Advancing land surface model development with satellite-based Earth observations. *Hydrology and Earth System Sciences*, 21(5):2483–2495.

Orth, R., Zscheischler, J., and Seneviratne, S. I. (2016). Record dry summer in 2015 challenges precipitation projections in Central Europe. *Scientific Reports*, 6:28334.

Poumadère, M., Mays, C., Mer, S. L., and Blong, R. (2005). The 2003 Heat Wave in France: Dangerous Climate Change Here and Now. *Risk Analysis*, 25(6):1483–1494.

Quesada, B., Vautard, R., Yiou, P., Hirschi, M., and Seneviratne, S. I. (2012). Asymmetric European summer heat predictability from wet and dry southern winters and springs. *Nature Climate Change*, 2(10):736–741.

Rebetez, M., Dupont, O., and Giroud, M. (2009). An analysis of the July 2006 heatwave extent in Europe compared to the record year of 2003. *Theoretical and Applied Climatology*, 95(1-2):1–7.

Robine, J.-M., Cheung, S. L. K., Le Roy, S., Van Oyen, H., Griffiths, C., Michel, J.-P., and Herrmann, F. R. (2008). Death toll exceeded 70,000 in Europe during the summer of 2003. *Comptes Rendus Biologies*, 331(2):171–178.

Russo, S., Sillmann, J., and Fischer, E. M. (2015). Top ten European heatwaves since 1950 and their occurrence in the coming decades. *Environmental Research Letters*, 10(12):124003.

Schrodin, R. and Heise, E. (2001). The Multi Layer Version of the DWD Soil Model TERRA_lm. *COSMO Technical Report*, 2:1–16.

Schulz, J.-P. and Vogel, G. (2017). An improved representation of the surface temperature including the effects of vegetation in the land surface scheme TERRA. http://www.cosmo-model.org/content/tasks/workGroups/wg3b/docs/TERRA_improvements_201703_jps.pdf. Last checked on May 30, 2015.

Schär, C. and Jendritzky, G. (2004). Climate change: Hot news from summer 2003. *Nature*, 432(7017):559–560.

Schär, C., Vidale, P. L., Lüthi, D., Frei, C., Häberli, C., Liniger, M. A., and Appenzeller, C. (2004). The role of increasing temperature variability in European summer heatwaves. *Nature*, 427(6972):332–336.

Sellers, P., Randall, D., Collatz, G., Berry, J., Field, C., Dazlich, D., Zhang, C., Collelo, G., and Bounoua, L. (1996). A Revised Land Surface Parameterization (SiB2) for Atmospheric GCMS. Part I: Model Formulation. *Journal of Climate*, 9(4):676–705.

Sellers, P. J., Dickinson, R. E., Randall, D. A., Betts, A. K., Hall, F. G., Berry, J. A., Collatz, G. J., Denning, A. S., Mooney, H. A., Nobre, C. A., Sato, N., Field, C. B., and Henderson-Sellers, A. (1997). Modeling the Exchanges of Energy, Water, and Carbon Between Continents and the Atmosphere. *Science*, 275(5299):502–509.

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., and Teuling, A. J. (2010). Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews*, 99(3-4):125–161.

Seneviratne, S. I., Donat, M. G., Pitman, A. J., Knutti, R., and Wilby, R. L. (2016). Allowable CO_2 emissions based on regional and impact-related climate targets. *Nature*, 529(7587):477–483.

Seneviratne, S. I., Lüthi, D., Litschi, M., and Schär, C. (2006). Land–atmosphere coupling and climate change in Europe. *Nature*, 443(7108):205–209.

Seneviratne, S. I., Wilhelm, M., Stanelle, T., Hurk, B., Hagemann, S., Berg, A., Cheruy, F., Higgins, M. E., Meier, A., Brovkin, V., Cluassen, M., Ducharne, A., Dufresne, J.-L., Findell, K. L., Ghattas, J., Lawrence, D. M., Malyshev, S., Rummukainen, M., and Smith, B. (2013). Impact of soil moisture-climate feedbacks on CMIP5 projections: First results from the GLACE-CMIP5 experiment. *Geophysical Research Letters*, 40(19):5212–5217.

Sippel, S., Otto, F. E. L., Flach, M., and van Oldenborgh, G. J. (2016). The Role of Anthropogenic Warming in 2015 Central European Heat Waves. *Bulletin of the American Meteorological Society*, 97(12):S51–S56.

Smiatek, G., Rockel, B., and Schättler, U. (2008). Time invariant data preprocessor for the climate version of the COSMO model (COSMO-CLM). *Meteorologische Zeitschrift*, 17:395–405.

Tanre, D., Geleyn, J., and Slingo, J. (1984). First Results of the Introduction of an Advanced Aerosol - Radiation Interaction in the ECMWF Low Resolution Global Model. *Aerosol and their Climatic Effects*, pages 133–177.

Tegen, I., Hollrig, P., Chin, M., Fung, I., Jacob, D., and Penner, J. (1997). Contribution of different aerosol species to the global aerosol extinction optical thickness: Estimates from model results. *Journal of Geophysical Research: Atmospheres*, 102(D20):23895–23915.

Trigo, I. F., Boussetta, S., Viterbo, P., Balsamo, G., Beljaars, A., and Sandu, I. (2015). Comparison of model land skin temperature with remotely sensed estimates and assessment of surface-atmosphere coupling: MODEL SKIN TEMPERATURE AND SATELLITE LST. *Journal of Geophysical Research: Atmospheres*, 120(23):12,096–12,111.

Vidale, P. L., Lüthi, D., Wegmann, R., and Schär, C. (2007). European summer climate variability in a heterogeneous multi-model ensemble. *Climatic Change*, 81(1):209–232.

Wang, A., Barlage, M., Zeng, X., and Draper, C. S. (2014). Comparison of land skin temperature from a land model, remote sensing, and in situ measurement: Comparison of land skin temperature. *Journal of Geophysical Research: Atmospheres*, 119(6):3093–3106.

Wielicki, B. A., Barkstrom, B. R., Harrison, E. F., Lee, R. B., Smith, G. L., and Cooper, J. E. (1996). Clouds and the Earth's Radiant Energy System (CERES): An Earth Observing System Experiment. *Bulletin of the American Meteorological Society*, 77(5):853–868.

Wouters, H., De Ridder, K., Poelmans, L., Willems, P., Brouwers, J., Hosseinzadehtalaei, P., Tabari, H., Vanden Broucke, S., van Lipzig, N. P. M., and Demuzere, M. (2017). Heat stress increase under climate change twice as large in cities as in rural areas: A study for a densely populated midlatitude maritime region. *Geophysical Research Letters*, 44(17):2017GL074889.

Zheng, W., Wei, H., Wang, Z., Zeng, X., Meng, J., Ek, M., Mitchell, K., and Derber, J. (2012). Improvement of daytime land surface skin temperature over arid regions in the NCEP GFS model and its impact on satellite data assimilation: LST IMPROVEMENT AND DATA ASSIMILATION. *Journal of Geophysical Research: Atmospheres*, 117(D6):1–14.

Ziv, Y. (2017). COSMO Priority Task: TERRA Nova.

Zubler, E. M., Lohmann, U., Lüthi, D., and Schär, C. (2011). Intercomparison of aerosol climatologies for use in a regional climate model over Europe. *Geophysical Research Letters*, 38(15).

6 Appendix

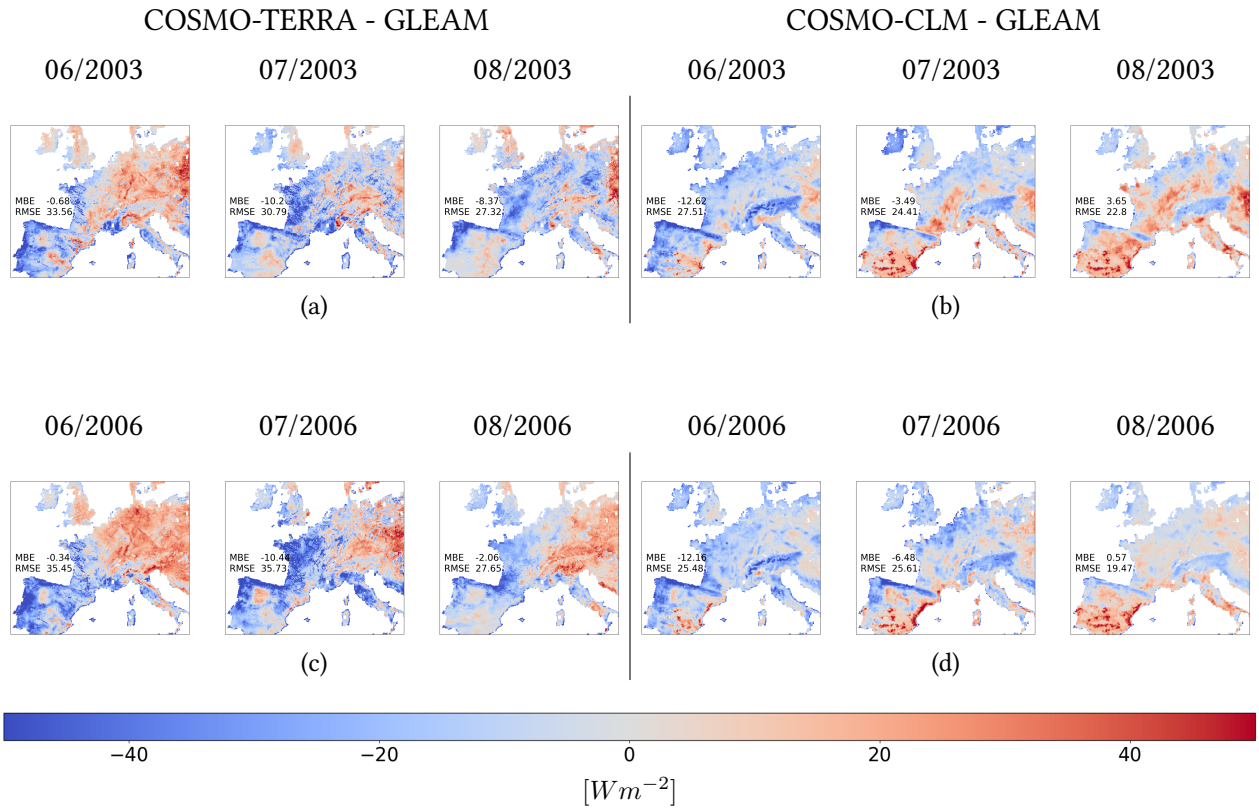


Figure A1: Monthly mean bias (model - observations) in daily mean latent heat flux for COSMO-TERRA (a, c) and COSMO-CLM (b, d) as compared to the GLEAM dataset for JJA 2003 (a,b) and 2006 (c, d). Positive values indicate overestimation of the flux in models. RMSE and MBE for each month are printed on the maps. Data where no measurements are available in the observations are excluded.

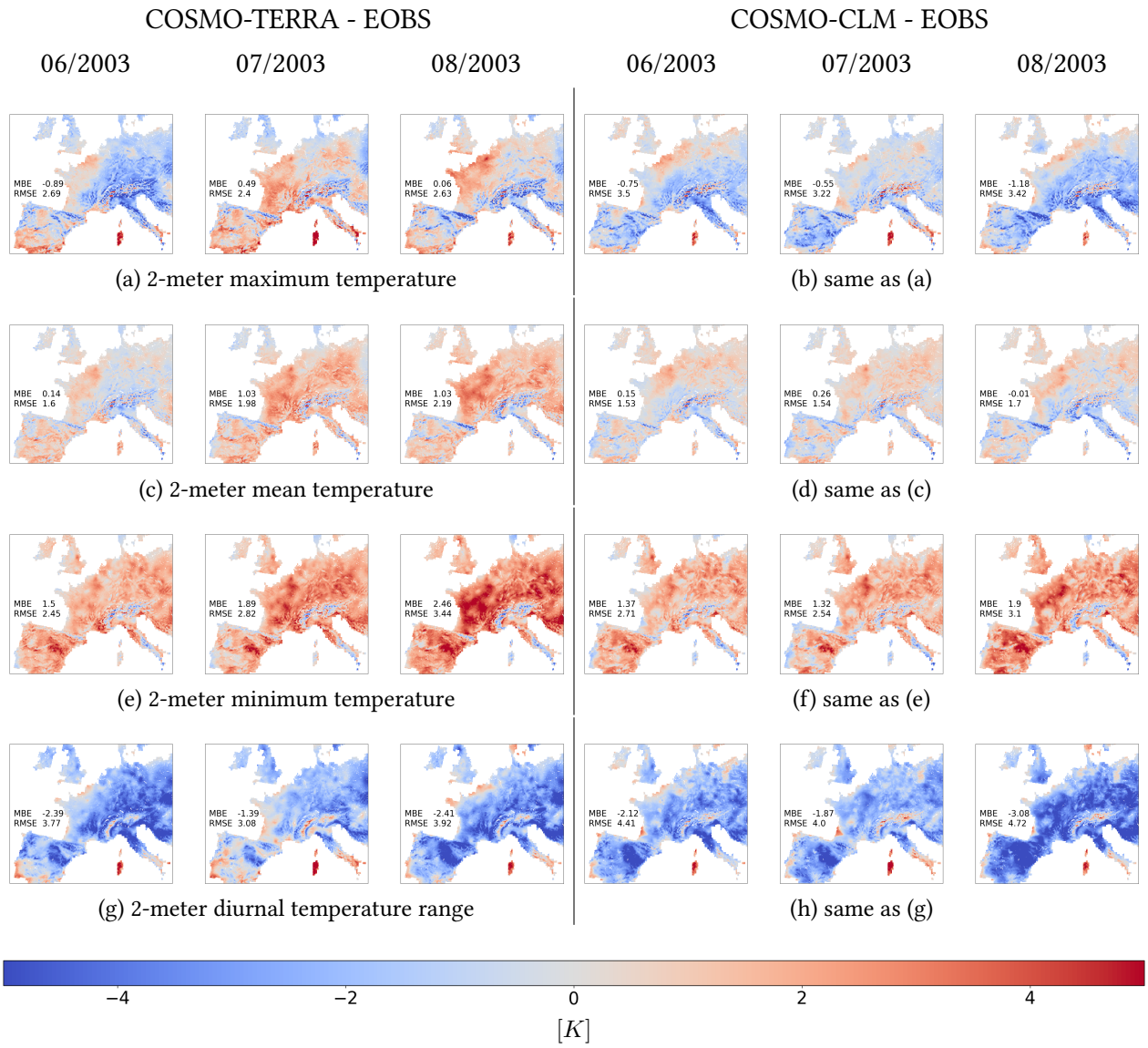


Figure A2: Monthly mean bias (model - observations) in (a, b) daily maximum 2-meter temperature, (c, d) daily mean 2-meter temperature, (e, f) daily minimum 2-meter temperature and (g, h) diurnal 2-meter temperature range for COSMO-TERRA (a, c, e, g) and COSMO-CLM (b, d, f, e) as compared to the EOBS dataset for JJA 2003. RMSE and MBE for each month are printed on the maps. Data where no measurements are available in EOBS are excluded.

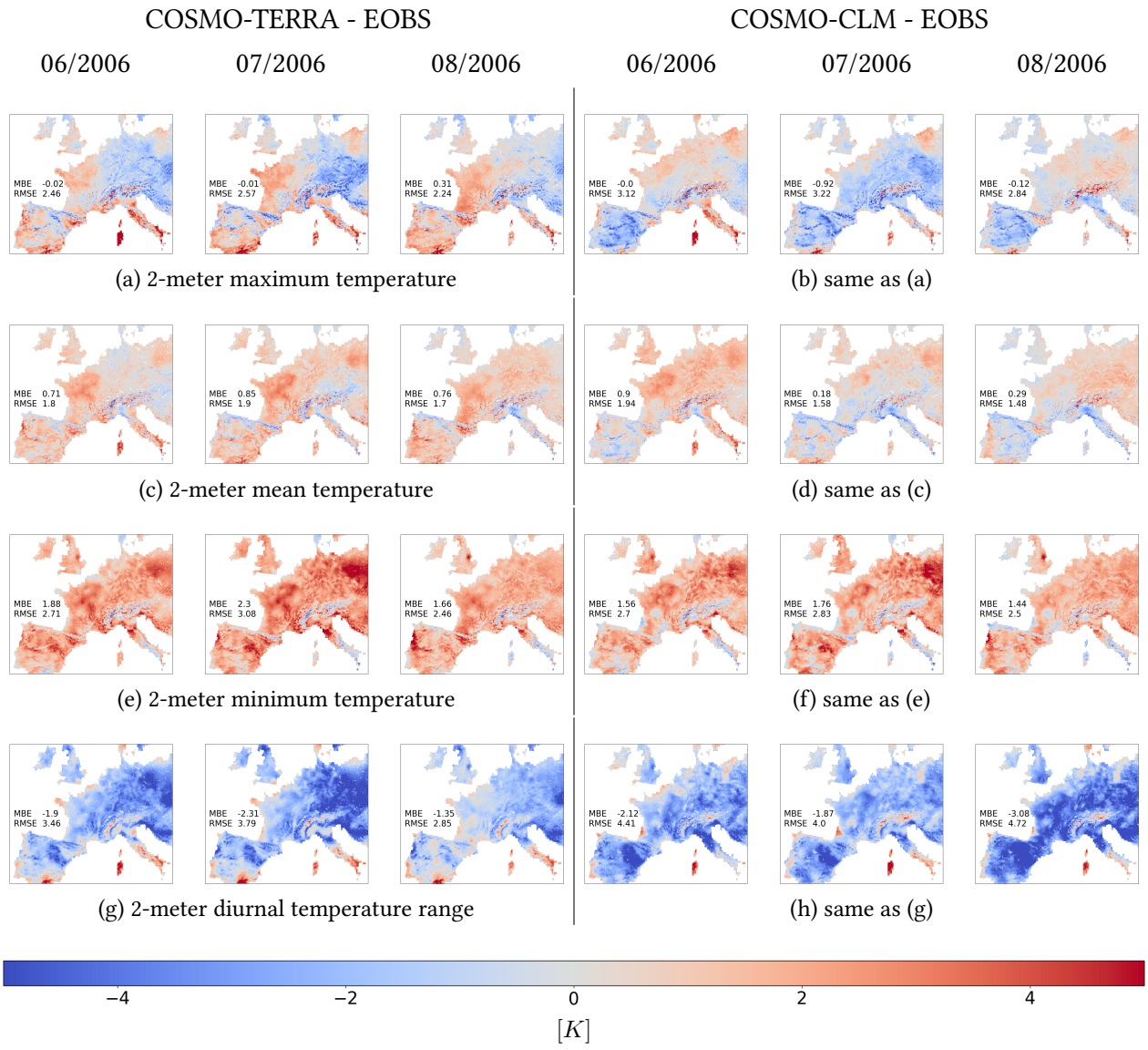


Figure A3: Monthly mean bias (model - observations) in (a, b) daily maximum 2-meter temperature, (c, d) daily mean 2-meter temperature, (e, f) daily minimum 2-meter temperature and (g, h) diurnal 2-meter temperature range for COSMO-TERRA (a, c, e, g) and COSMO-CLM (b, d, f, e) as compared to the EOBS dataset for JJA 2006. RMSE and MBE for each month are printed on the maps. Data where no measurements are available in EOBS are excluded.

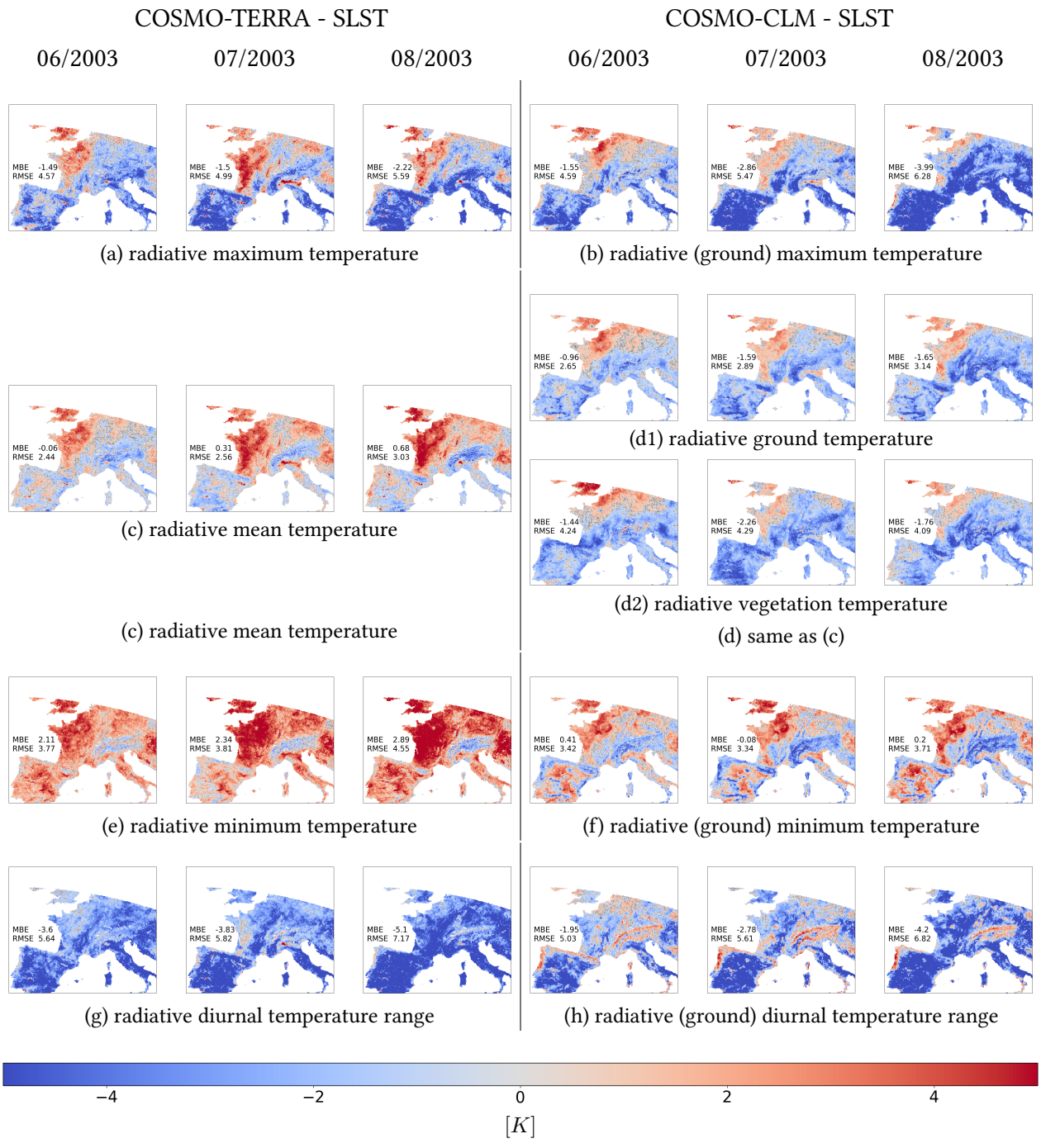


Figure A4: Monthly mean bias (model - observations) in (a, b) daily maximum radiative temperature, (c, d) daily mean radiative temperature, (e, f) daily minimum radiative temperature and (g, h) diurnal radiative temperature range for COSMO-TERRA (a, c, e, g) and COSMO-CLM (b, d, f, h) as compared to the SLST dataset for JJA 2003. For evaluation with COSMO-CLM, the CLM variable ground temperature is used in all plots except (d2), where vegetation temperature is shown additionally for comparison (see Section 2.2 for distinction). Pixels are stippled where the uncertainty of the satellite measurement is larger than the reported bias at this point. RMSE and MBE for each month are printed on the maps. Data where no measurements are available in SLST are excluded. Note that this especially includes all data points where cloud cover was detected by the satellite.

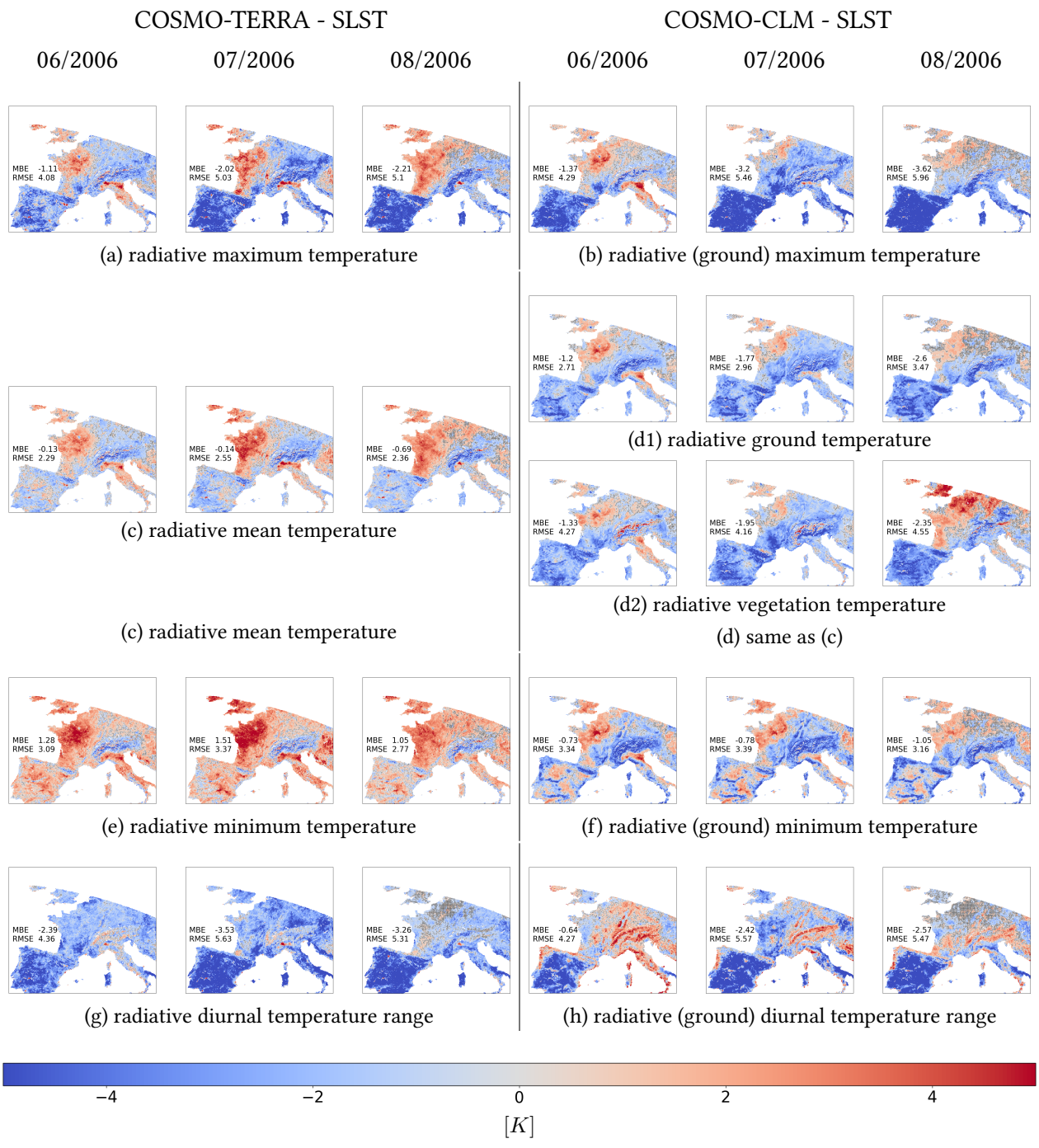


Figure A5: Monthly mean bias (model - observations) in (a, b) daily maximum radiative temperature, (c, d) daily mean radiative temperature, (e, f) daily minimum radiative temperature and (g, h) diurnal radiative temperature range for COSMO-TERRA (a, c, e, g) and COSMO-CLM (b, d, f, h) as compared to the SLST dataset for JJA 2006. For evaluation with COSMO-CLM, the CLM variable ground temperature is used in all plots except (d2), where vegetation temperature is shown additionally for comparison (see Section 2.2 for distinction). Pixels are stippled where the uncertainty of the satellite measurement is larger than the reported bias at this point. RMSE and MBE for each month are printed on the maps. Data where no measurements are available in SLST are excluded. Note that this especially includes all data points where cloud cover was detected by the satellite.

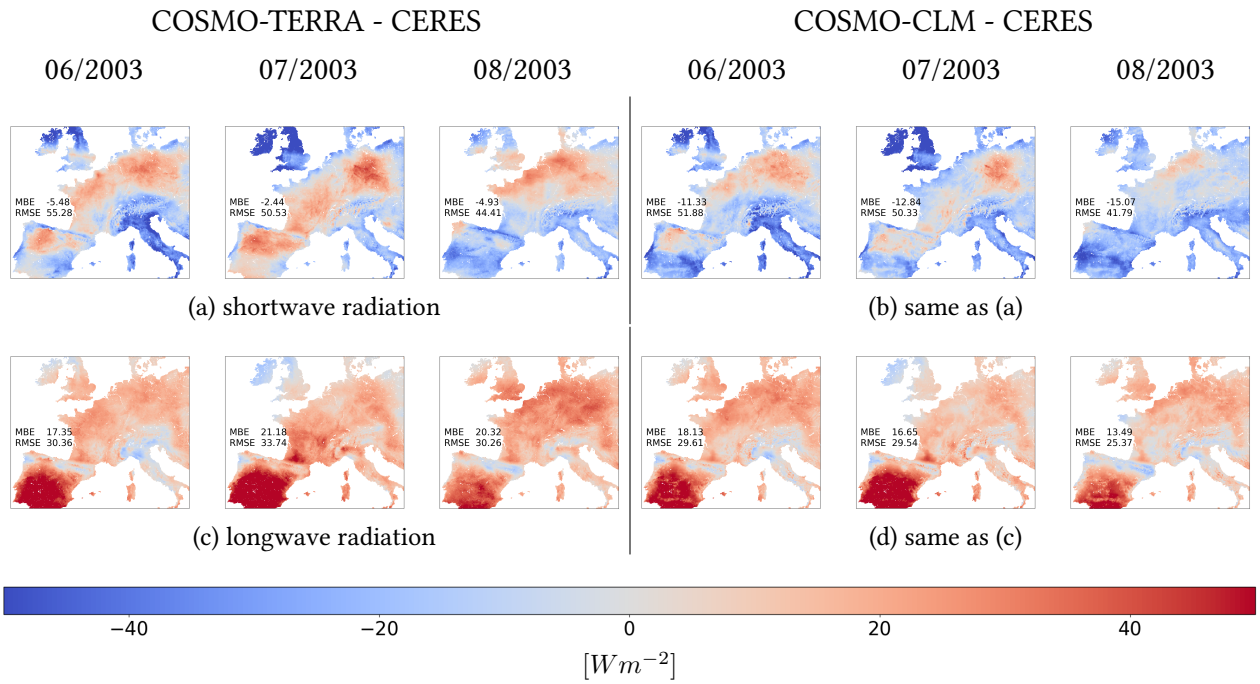


Figure A6: Monthly mean bias (model - observations) in daily mean (a, b) shortwave radiation and (c, d) longwave radiation for COSMO-TERRA (a, c) and COSMO-CLM (b, d) as compared to the CERES dataset for JJA 2003. RMSE and MBE for each month are printed on the maps. Data where no measurements are available in the observations are excluded.

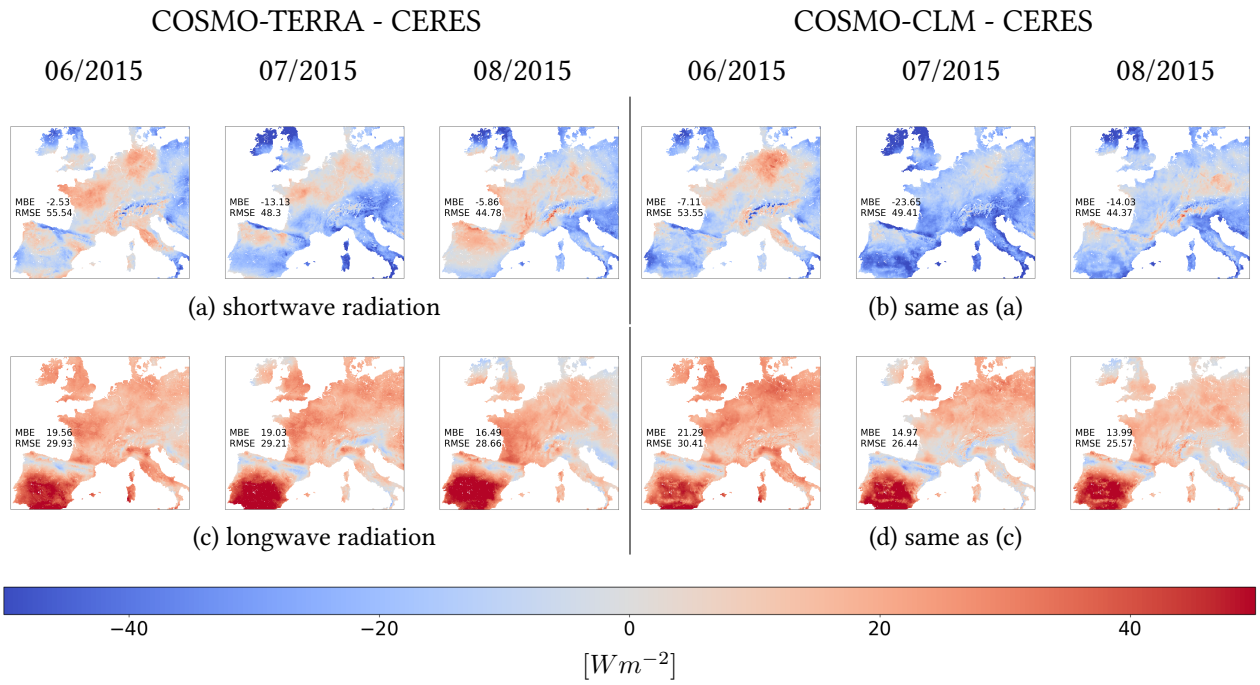


Figure A7: Monthly mean bias (model - observations) in daily mean (a, b) shortwave radiation and (c, d) longwave radiation for COSMO-TERRA (a, c) and COSMO-CLM (b, d) as compared to the CERES dataset for JJA 2006. RMSE and MBE for each month are printed on the maps. Data where no measurements are available in the observations are excluded.

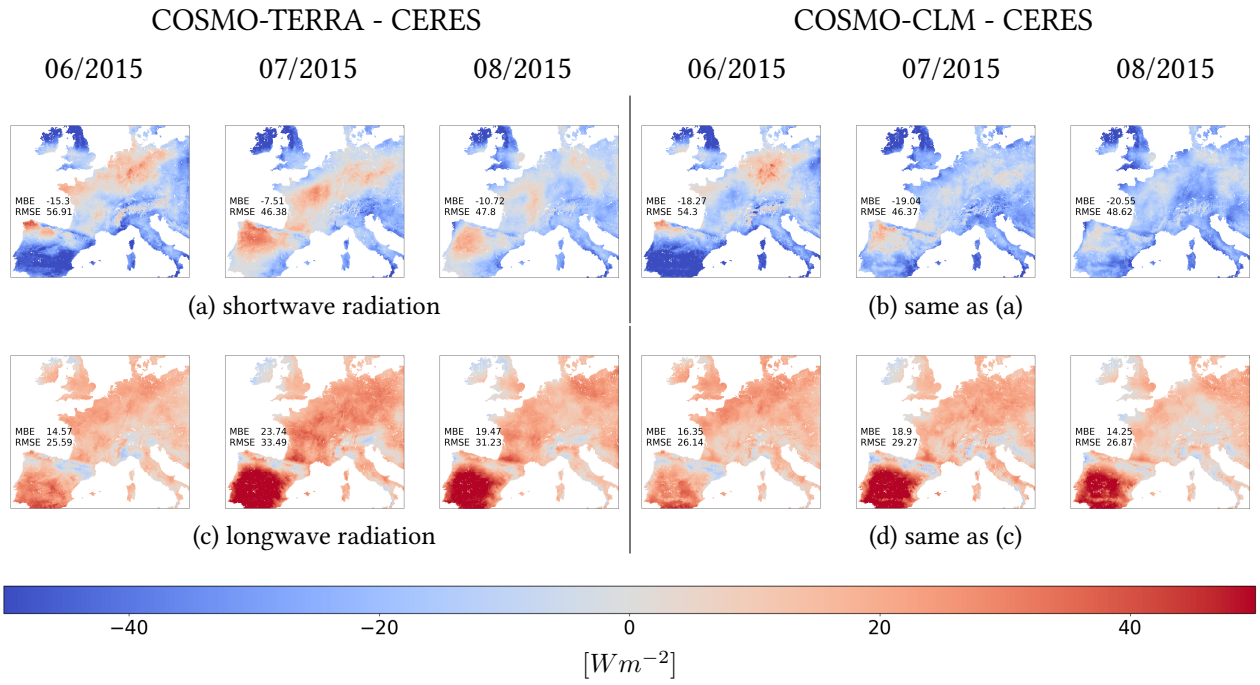


Figure A8: Monthly mean bias (model - observations) in daily mean (a, b) shortwave radiation and (c, d) longwave radiation for COSMO-TERRA (a, c) and COSMO-CLM (b, d) as compared to the CERES dataset for JJA 2015. RMSE and MBE for each month are printed on the maps. Data where no measurements are available in the observations are excluded.

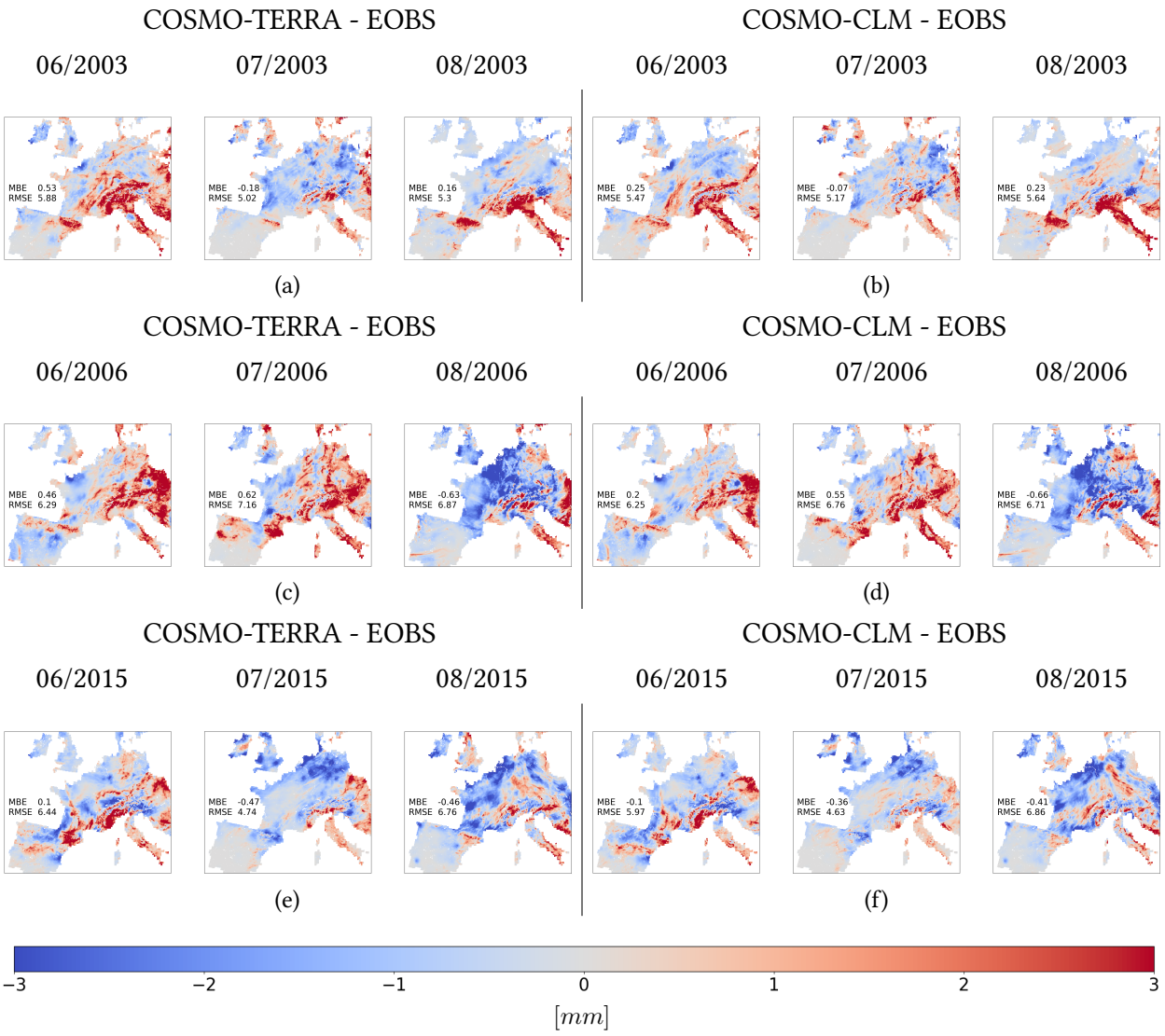
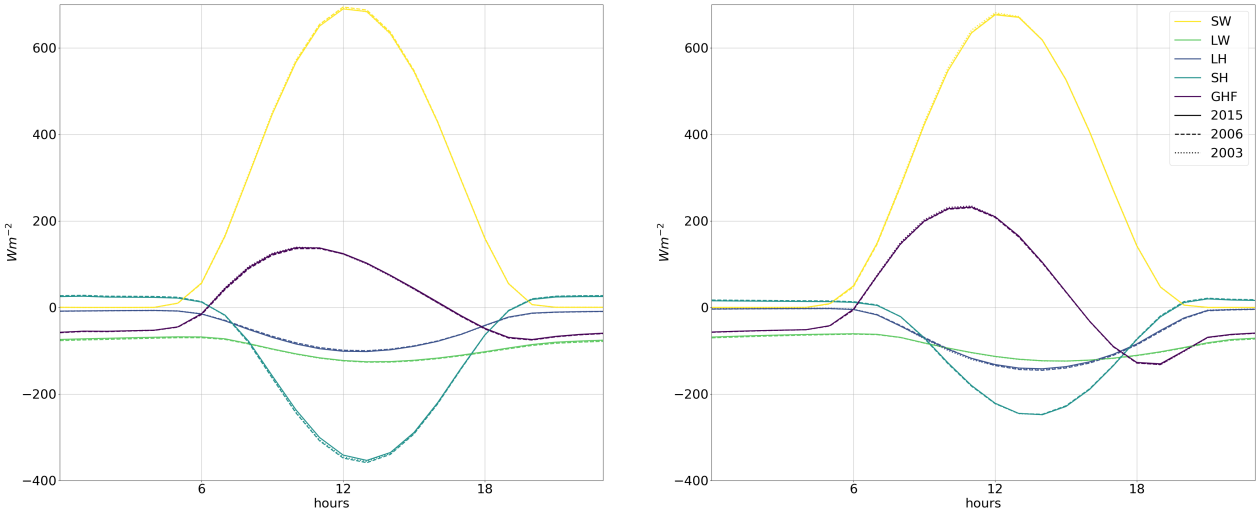
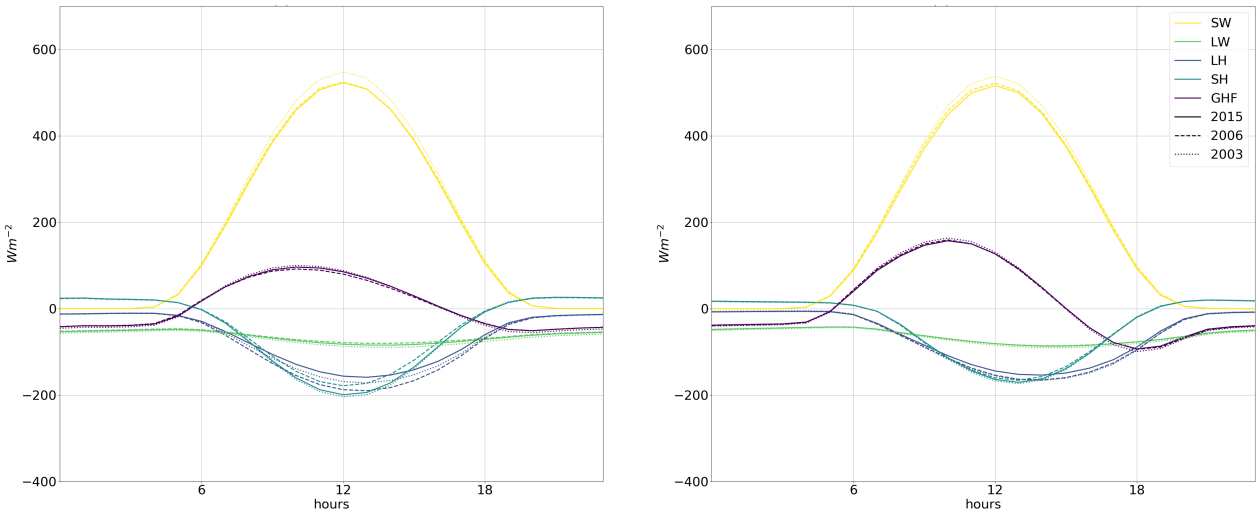


Figure A9: Monthly mean bias (model - observations) in daily mean precipitation for COSMO-TERRA (a, c) and COSMO-CLM (b, d) as compared to the CERES dataset for JJA 2003 (a, b) and 2006 (c, d). RMSE and MBE for each month are printed on the maps. Data where no measurements are available in the observations are excluded.



(a) Southern European domain



(b) Northern European domain

Figure A10: Mean diurnal cycle of net shortwave radiation (SW), net longwave radiation (LW), net latent heat (LH) and net sensible heat (SH) at surface for COSMO-TERRA (left panels) and COSMO-CLM (right panels) averaged over the (a) Southern European and (b) Northern European domain for JJA in 2003, 2006, 2015. Note that in agreement with the COSMO convention, fluxes towards the surface are defined positive.

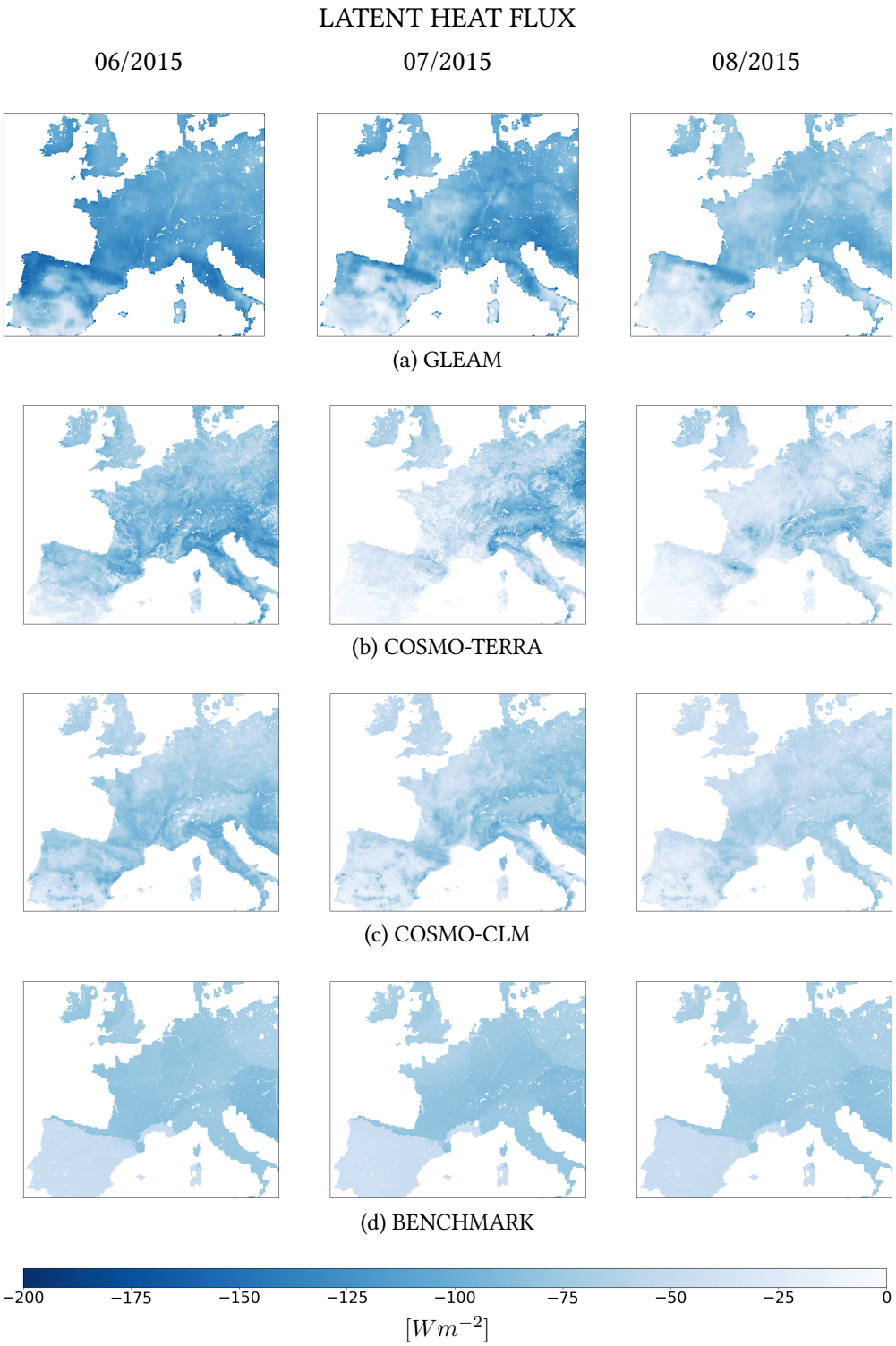


Figure A11: Monthly mean latent heat flux in (a) GLEAM, (b) COSMO-TERRA, (c) COSMO-CLM and (d) the benchmark for JJA 2015. Data where no measurements are available in observations are excluded.

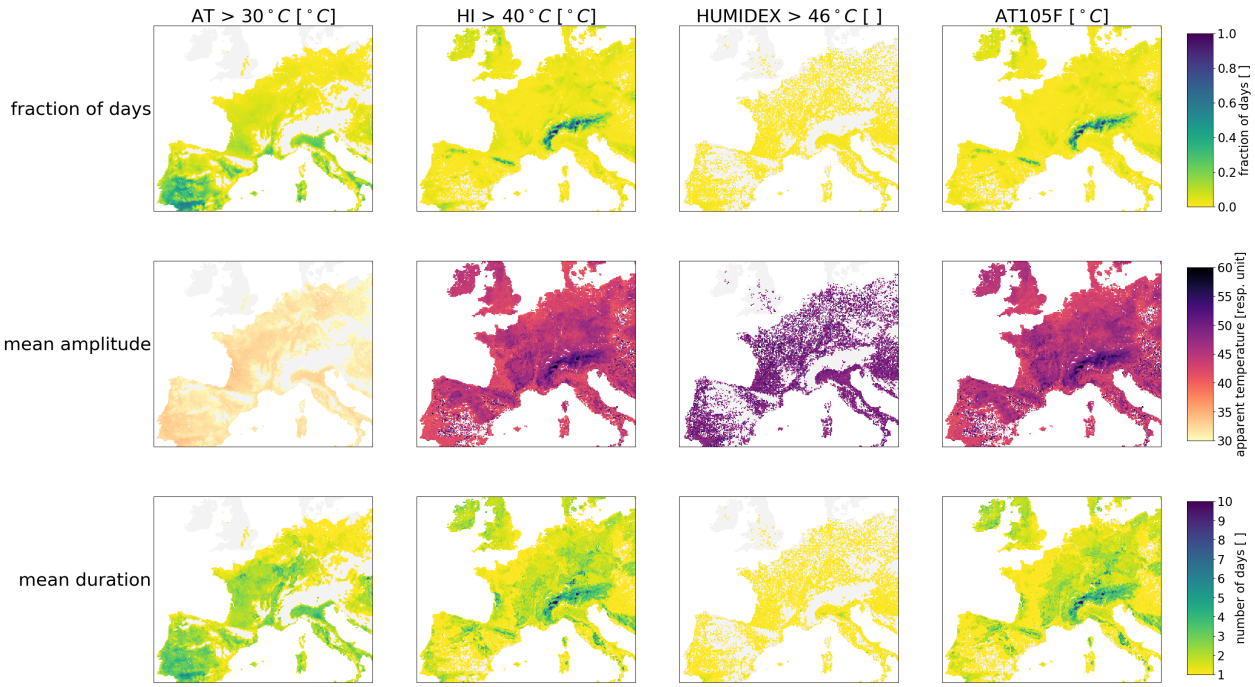
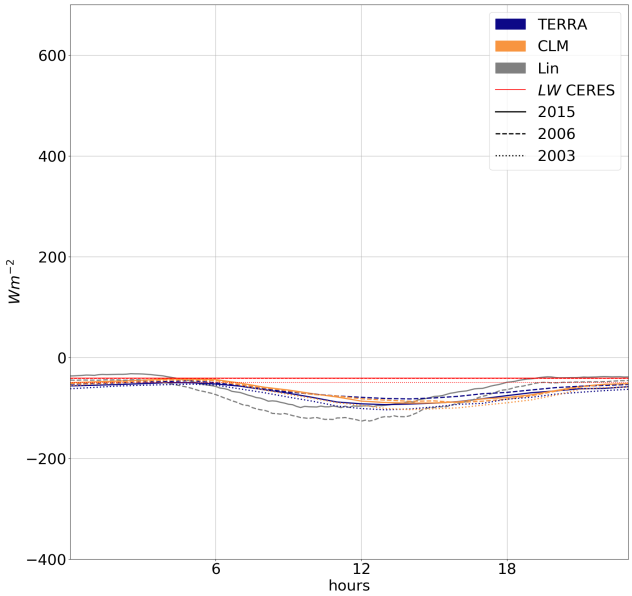


Figure A12: Magnitude of heat extremes exceeding human comfort thresholds over Europe for all three summers expressed as fraction of days affected (upper panels), mean amplitude of heat extremes (middle panels) and mean duration of heat extremes (lower panels) for COSMO-CLM. Indices used here are, from left to right, apparent temperature after Buzan et al. (2015), Heat Index, HUMIDEX and AT105F (see Fischer and Schär, 2010).



(a) Longwave radiation

Figure A13: Comparison of diurnal cycle of longwave radiation (Supplementary Figure A13) at the Lindenberg station and the two models at the associated model pixel averaged over the summer months for the three summers. The CERES daily estimate is included (red line).

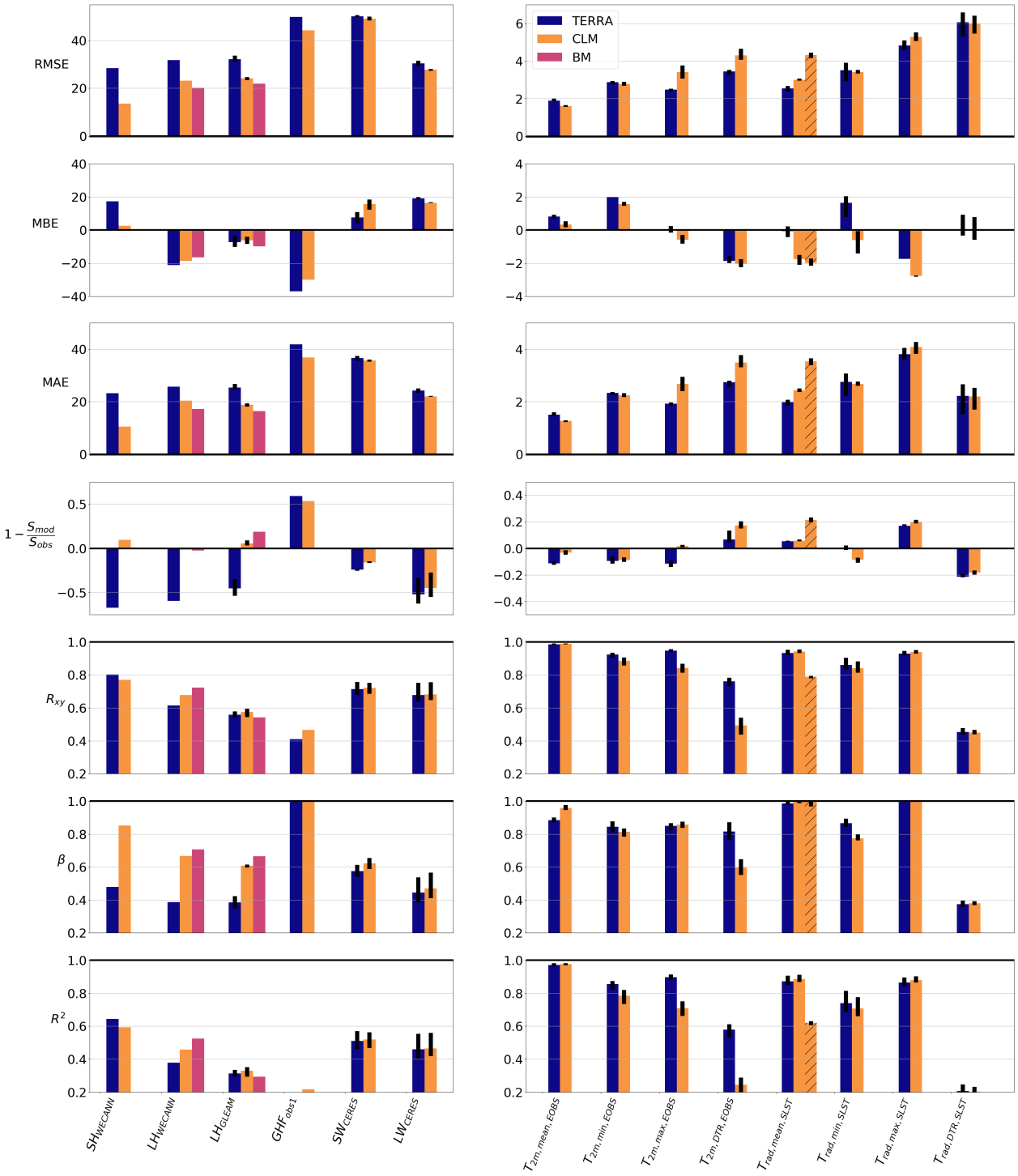


Figure A14: Performance metrics with all evaluation datasets available. Measures are calculated for daily values in JJA of 2015, 2006 and 2003, except for the benchmark. For $T_{rad,mean,SLST}$, hatched bars are comparison to T_v and unhatched bars for T_{rad} in CLM. For WECANN, only 2015 is available. Black thick lines indicate the best value of each measure. Errorbars indicate the difference between the two years. For similar analysis in climate mode see Supplementary Figure A17.

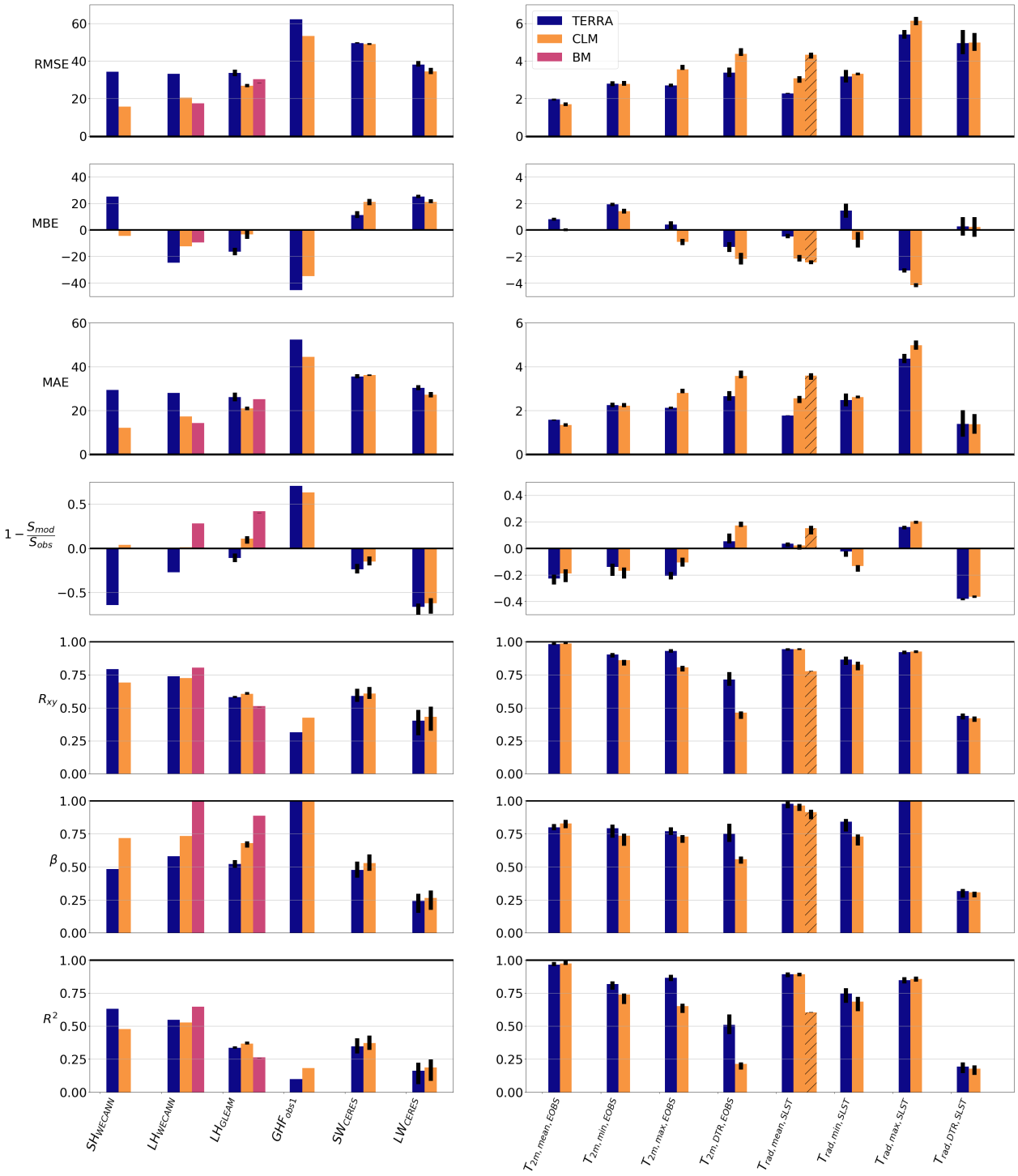


Figure A15: Performance metrics with all evaluation datasets available for the Southern European domain. Measures are calculated for daily values in JJA of 2015 and 2003. For $T_{rad,mean,SLST}$, hatched bars are comparison to T_v and unhatched bars for T_{rad} in CLM. For WECANN, only 2015 is available. Black thick lines indicate the best value of each measure. Errorbars indicate the difference between the two years. For similar analysis in climate mode see Supplementary Figure A17.

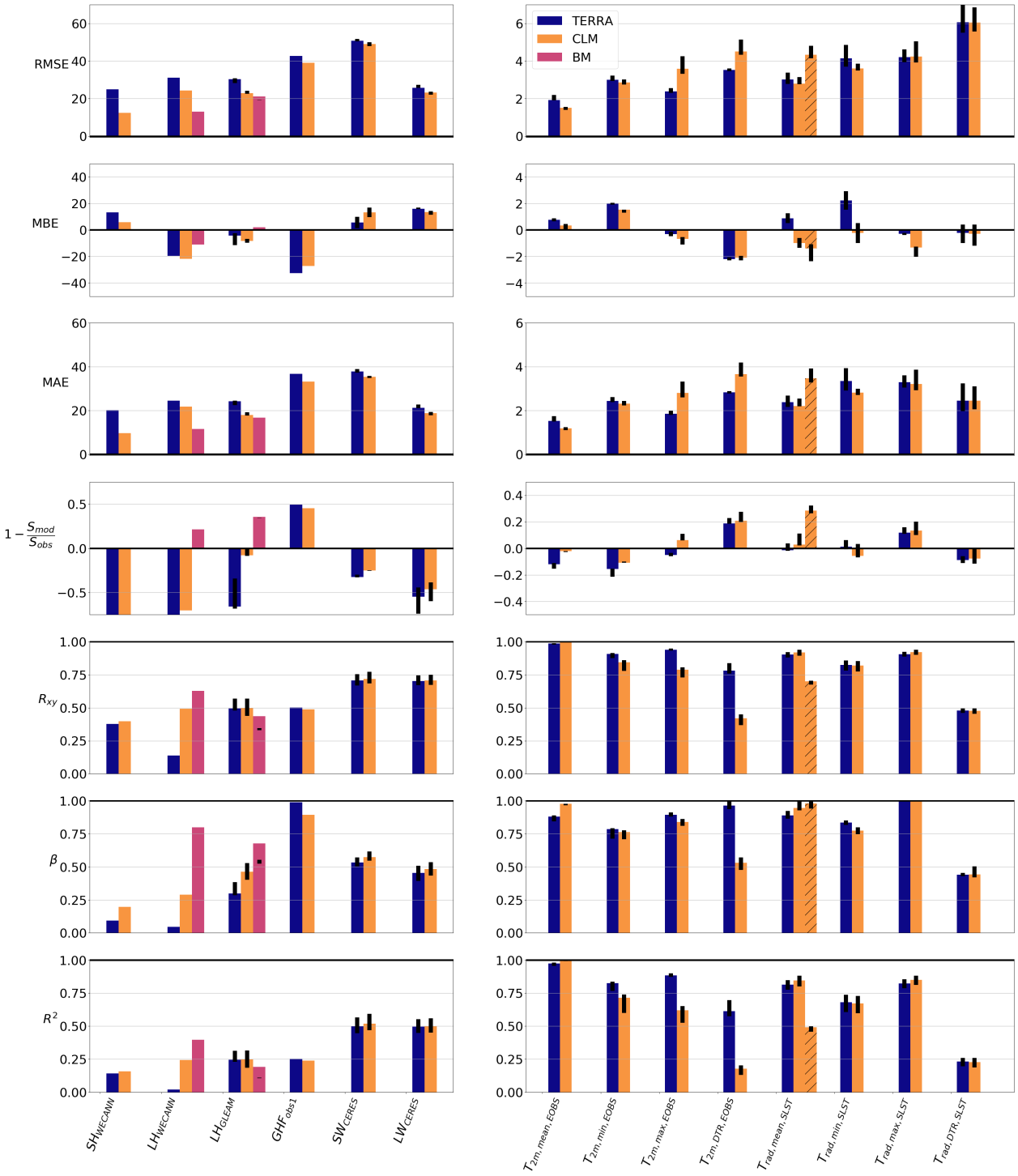


Figure A16: Performance metrics with all evaluation datasets available for the Northern European domain. Measures are calculated for daily values in JJA of 2015 and 2003. For $T_{rad,mean,SLST}$, hatched bars are comparison to T_v and unhatched bars for T_{rad} in CLM. For WECANN, only 2015 is available. Black thick lines indicate the best value of each measure. Errorbars indicate the difference between the two years. For similar analysis in climate mode see Supplementary Figure A17.

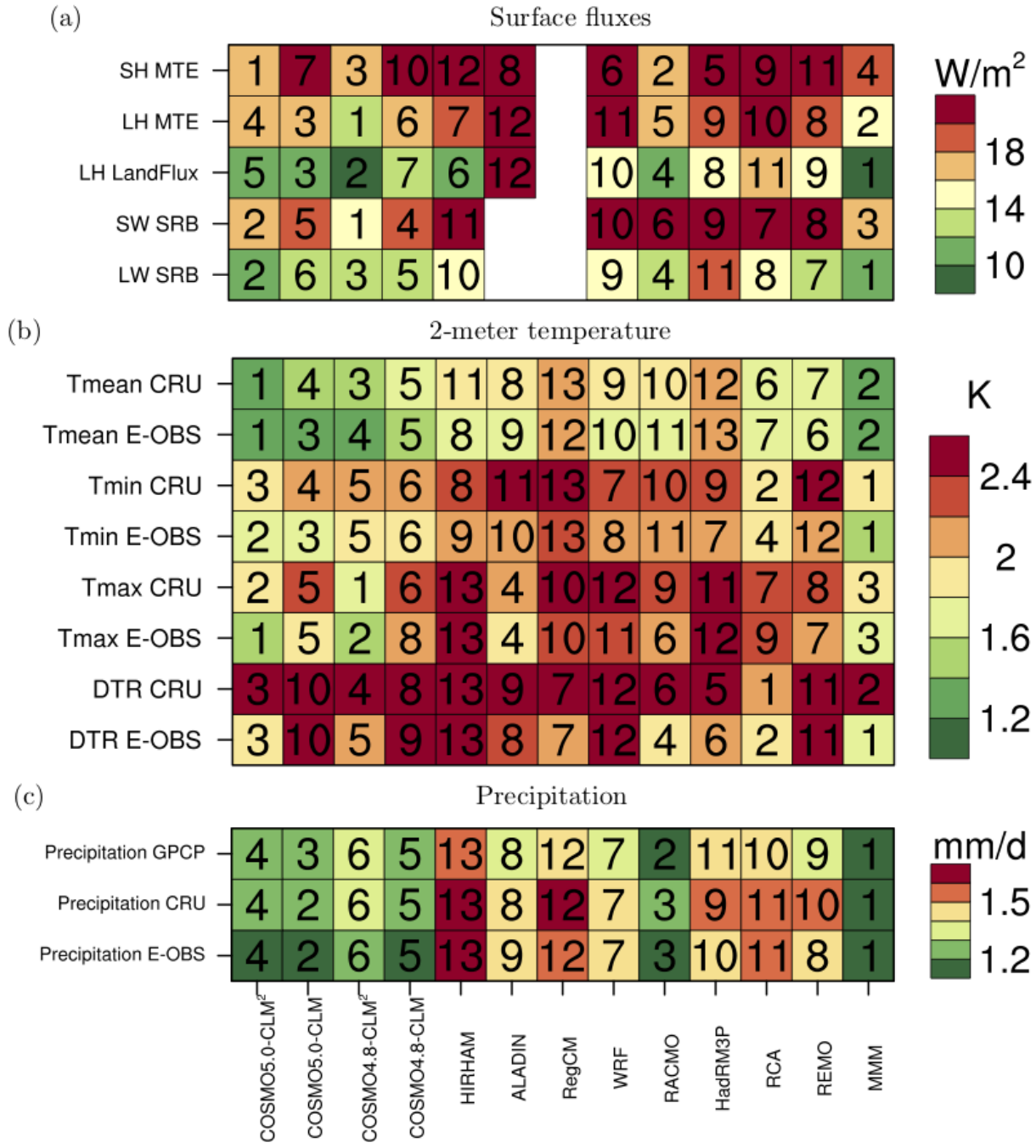


Figure A17: Results from (Davin et al., 2016) additionally including the COSMO version 5.0 used in this study. RMSE of model variables with available datasets over over Europe (-10W - 30E; 36N - 70N) from monthly values over multiple years. MMM denotes the multi-model mean of all EURO-CORDEX models excluding COSMO-CLM. For more information see also (Davin et al., 2016) Figure 1. Figure by Edouard Davin.

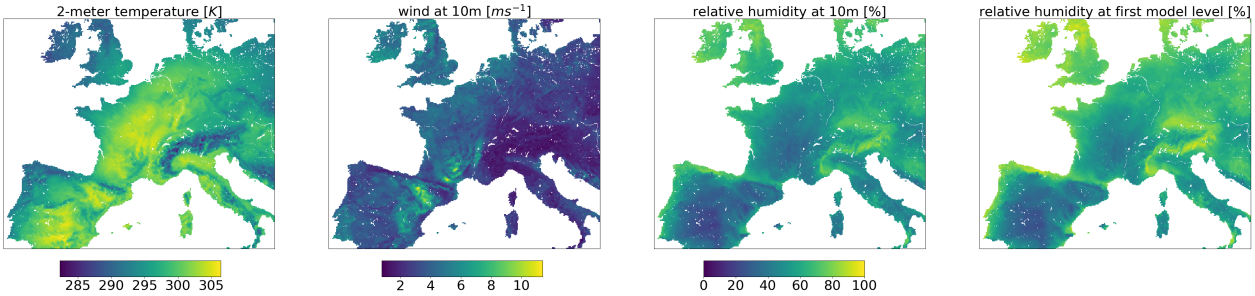


Figure A18: Average (a) temperature, (b) wind speed and relative humidity at 2 m (c) and at the first model level (d) averaged over JJA 2015 for COSMO-TERRA over the whole domain as included in the human comfort indices in Figure 17.

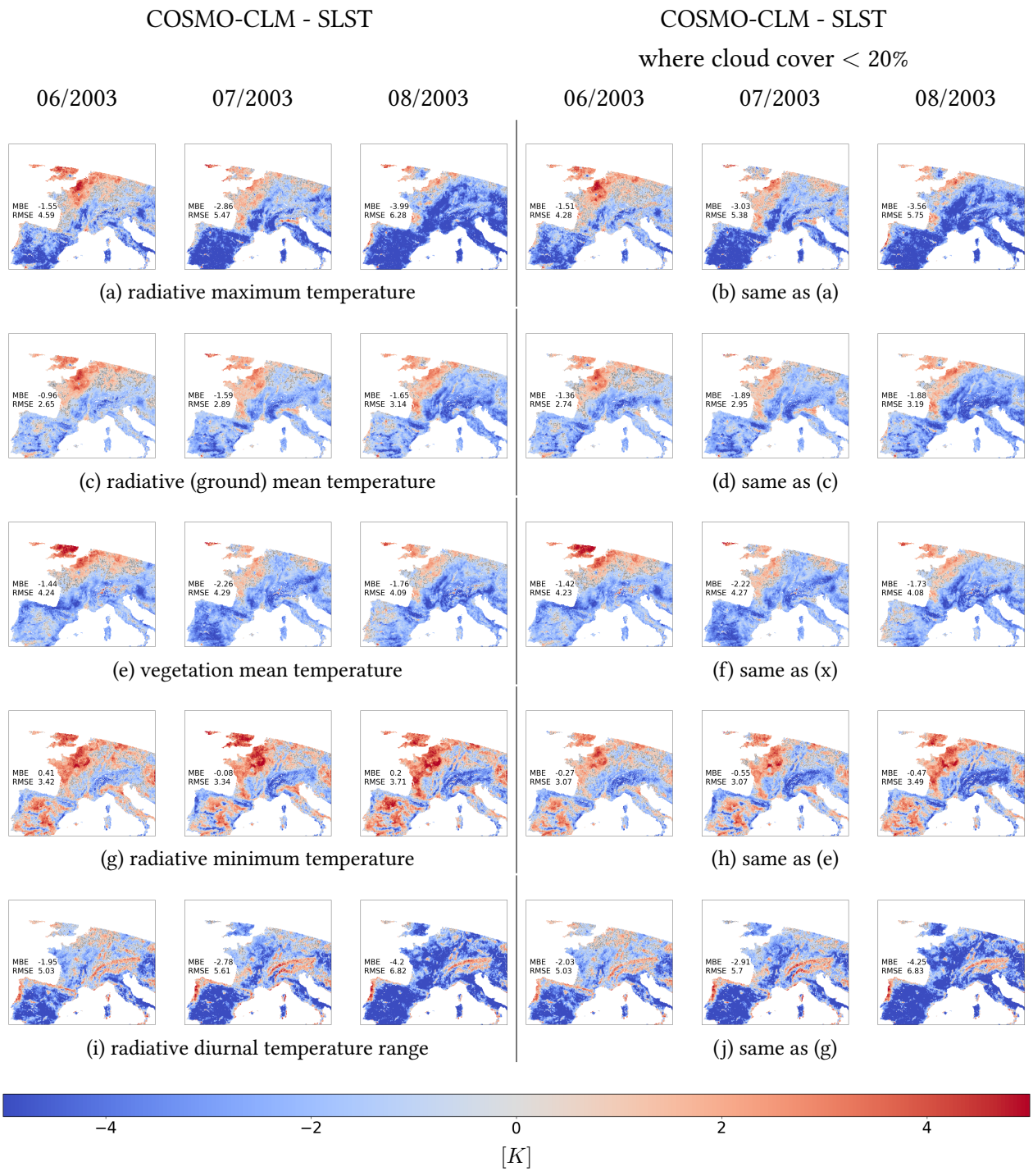


Figure A19: Comparison of 2003 bias maps of CLM vs SLST (CLM - SLST) for (a, c, e, g, i, i.e. left panels) all cloud cover fractions in model (same as Figure 12 for 2003) and (b, d, f, h, j, i.e. right panels) only for points where modelled cloud cover is below 20%. Note that cloud cover is only available for 2003 CLM runs, hence a threshold for modelled cloud cover is not applied throughout this study. This plot is merely to show that the difference, if applied to all runs, is likely small.