Deep learning for quality control of surface physiographic fields using satellite Earth observations

Tom Kimpson[1], Margarita Choulga[2], Matthew Chantry[2], Gianpaolo Balsamo[2], Souhail Boussetta[2], Peter Dueben[2], and Tim Palmer[1]

[1]Department of Physics, University of Oxford, Oxford, UK
[2]Research Department, European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK

Correspondence: Tom Kimpson (tom.kimpson@physics.ox.ac.uk)

Abstract. About 2/3 of all densely populated areas (i.e. at least 300 inhabitants per km^2) around the globe are situated within a 9 km radius of a permanent waterbody (i.e. inland water or sea/ocean coast), since inland water sustains the vast majority of human activities. Water bodies exchange mass and energy with the atmosphere and need to be accurately simulated in numerical weather prediction and climate modelling as they strongly influence the lower boundary conditions such as skin temperatures, turbulent latent and sensible heat fluxes and moisture availability near the surface. All the non-ocean water (resolved and sub-grid lakes and coastal waters) are represented in the Integrated Forecasting System (IFS) of the European Centre for Medium-Range Weather Forecasts (ECMWF) model, by the Fresh-water Lake (FLake) parametrisation, which treats $\sim$ 1/3 of the land. It is a continuous enterprise to update the surface parametrization schemes and their input fields to better represent small-scale processes. It is, however, difficult to quickly determine both the accuracy of an updated parametrisation, and the added value gained for the purposes of numerical modelling. The aim of our work is to quickly and automatically assess the benefits of an updated lake parametrisation making use of a neural network regression model trained to simulate satellite observed surface skin temperatures. In order to rapidly assess accuracy of new surface physiography fields and if their use in the model increase predictability a neural network regression model (further in the text referred as VESPER) that can learn the mapping between a set of features x and targets y is constructed. In this case the features are the atmospheric and surface model fields (such as 2 metre temperature from ERA5 reanalysis) and the surface physiographic fields (such as orography and vegetation cover used to produce ERA5 reanalysis), see Table 1 for the full list of variables used. The target is the satellite land surface temperature (skin temperature from MODIS Aqua Day MYD11A1 v006 collection). VESPER can then make predictions about the skin temperature given a set of input variables (i.e. atmospheric and surface model fields, and surface physiographic fields). In turn, these predictions can then be compared against observations (i.e. satellite skin temperature). We deploy this tool to determine the accuracy of recent upgrades to the FLake parametrisation, namely the improved permanent lake cover and the capacity to represent seasonally varying water bodies (i.e. ephemeral lakes). We show that for grid-cells where the lake fields have been updated, the prediction accuracy in the land surface temperature improves by 0.45 K on average, whilst for the subset of points where the lakes have been exchanged for bare ground (or vice versa) the improvement is 1.12 K. We also show that updates to the glacier cover improve further the prediction accuracy by 0.14 K. The inclusion of seasonal water is shown to be particularly effective for grid points which are highly time variable, generally improving the simulation accuracy

by $\sim$1 K. The neural network regression model has proven to be useful and easily adaptable to assess unforeseen impacts of ancillary datasets, also detecting inappropriate changes of high vegetation to bare ground, which would lead to decreased the skin temperature simulation accuracy by 0.49 K, proving to be a valuable support to model development.

1 Introduction

Globally, there are $\sim$ 117 million lakes - defined as inland water bodies without lateral movement of water - making up around 3.7% of the Earth's land surface (Verpoorter et al., 2014). Their distribution is highly anisotropic, with the majority of lakes located between $45-75^{\circ}$N in the Boreal and Arctic regions. Lakes are highly important from the perspective of both numerical weather prediction and climate modelling as part of the EC-Earth model. For the latter, lakes generally influence the global carbon cycle as both sinks and sources of greenhouse gases; the majority of lakes are net heterotrophic, over saturated with CO_2 as a result of in lake respiration and so emit carbon into the atmosphere (Pace and Prairie, 2005; Tranvik et al., 2009). Total CO_2 emission from lakes is estimated at $1.25-2.30$ Pg of CO2-equivalents annually (DelSontro et al., 2018), nearly 20% of global CO_2 fossil fuel emissions, whilst lakes account for 9-24 $\%$ of CH_4 emissions, the second largest natural source after wetlands (Saunois et al., 2020). These rates of greenhouse gas emission are expected to rise further if the eutrophication of the Earth's lentic systems continues. With regards to weather, freezing and melting of the lake surface modifies the radiative and conductive properties and consequently affects the heat (latent, sensible) exchange and surface energy balance (Huang et al., 2019; lu et al., 2020; Franz et al., 2018). Considering particular examples, over Lake Victoria convective activity is suppressed during the day and peaks at night, leading to intense, hazardous thunderstorms (Thiery et al., 2015, 2017); Lake Ladoga can generate low level clouds which can cause variability in the 2m temperature of up to 10 K (Eerola et al., 2014); the Laurentian Great Lakes can cause intense winter snow storms (Vavrus et al., 2013) (Notaro et al., 2013). Moreover, as a result of the increased temperatures due to climate change, lakes become more numerous due to the melting of glaciers and permafrost. Additionally, the higher temperatures mean that previously permanent lake bodies become seasonal or intermittent. There is then evidently a huge potential return in the ability to accurately model the location, morphology and properties of lakes in weather and climate models.

The Integrated Forecasting System (IFS) at the European Centre for Medium Range Weather Forecasts (ECMWF) is used operationally for numerical weather prediction and climate modelling. Earth-system modelling in the IFS can be broadly categorised into large-scale and small-scale processes. Large-scale processes can be described by numerically solving the relevant set of differential equations, to determine e.g. the general circulation of atmosphere. Conversely, small-scale processes such as clouds or land-surface processes are represented via parametrisation. Accurate parametrisations are essential for the overall accuracy of the model. For example, the parametrisation of the land surface determines the sensible and latent heat fluxes, providing the lower boundary conditions for the equations of enthalpy and moisture in the atmosphere (Viterbo, 2002).

Lakes are incorporated in Earth-system models via parametrisation. At ECMWF the representation of lakes via parametrisation was first handled by introducing the Fresh water Lake model FLake (Mironov, 2008) into the IFS. FLake treats all resolved inland waterbodies (i.e. lakes, reservoirs, rivers which are dominating in a grid-cell) and unresolved or sub-grid water (i.e. small inland waterbodies and sea/ocean coastal waters which are present but not dominating in a grid-cell). Note that lake parameters are also an important part of the FLake model so when we refer in this work to “lake parametrisation" we mean both the model and the parameters. The broad impact of the FLake model and the important role that waterbodies play in human life can be illustrated by analysing ECMWF maps of the fractional land sea mask and the inland waterbody cover alongside maps of the population density (i.e. inhabitants per km^2) based on the population count for 2015 from the Global Human Settlement Layers (GHSL), Population Grid 1975-2030 (Schiavina et al., 2022; Freire et al., 2016) at 9 km horizontal resolution. Globally FLake is active over 11.1% of the grid-cells, with only 1.2% of them being resolved inland waters (i.e. water covers $\geq$50% of the grid-cell); considering only non-ocean (i.e. land) grid-cells, then FLake is active over 32.4% of the grid-cells with only 3.5% of them being resolved waters. According to the population data, only 4% of land is densely populated (i.e. at least 300 inhabitants per km^2); 64.5% of these areas being situated within a 9 km radius of a permanent waterbody (i.e. inland water or sea/ocean coast) with half of it (i.e. 31.2% of densely populated areas) being in the vicinity of at least 1 km^2 waterbody - emphasising how essential waterbodies are in human life. In some regions this role may be even more crucial than in the others. For example, only 2% of the North American region (similar for South American and North Asian regions) is densely populated with 45.7% (33.9% and 37.9% respectively) of the areas being in vicinity of at least 1 km^2 waterbody; for Europe even though it has more densely populated areas (16% of land is densely populated) still 37.4% of the population are in the vicinity of at least a 1 km^2 waterbody; for a rather dry continent like Africa only 5% of land is densely populated with 22.2% of these areas being close to at least a 1 km^2 waterbody; most striking in this sense is Australia where only 0.5 % of the land is populated, with two thirds of the population living within 9 km radius of a permanent waterbody of at least 1 km^2, with the majority of people living on the ocean coast.

It is a continuous enterprise to update the lake parametrization schemes and their input data fields to better represent small-scale surface processes. It is however challenging to accurately represent lakes in these parametrisations; the majority of lakes which are resolved at a 9km grid spacing have not had their morphology accurately measured, let alone monitored, whilst 28.9% of land and coastal cells are treated for sub-grid (i.e. covering half or less of a grid cell) water. When introducing an updated lake representation it is difficult apriori to determine the additional value gained through doing so. There are two key factors here:

- Are the updated fields closer to reality?
- Do the updated fields increase the accuracy of the model predictions?

The first point is straightforward; we want our parametrisation fields to better represent reality. If the lake depth of some lake is updated from 10m to 100m we want to be sure that 100m is closer to the true depth of the lake. For the second point, even if the updated fields are accurate, are they informative in the sense that they enable us to make more accurate predictions? For instance, the main target of lake parametrization is to reproduce lake surface water temperatures (and therefore evaporation

rates). If a lake parametrisation scheme is updated to better represent different types of inland waterbodies, the time variability of inland waterbodies and/or the lake morphology fields use more in situ measurements, does this additional information allow for more accurate predictions of the lake surface water temperatures? Is it therefore worthwhile to update the parametrisation in this way? Since the resulting updated fields are ultimately used operationally, it is essential to ensure the accuracy of the fields and prevent any potential degradation or instability of the model. This problem of quickly and automatically verifying the accuracy and information gain of updated lake parametrisations is the aim of this work.

Numerical weather prediction and climate modelling are fields that are inherently linked with large datasets and complex, non-linear interactions. It is therefore an area that is particularly well placed to benefit from the deployment of machine learning algorithms. At ECMWF, advanced machine learning techniques have been used for parametrisation emulation via neural networks (Chantry et al., 2021), 4D-Var data assimilation (Hatfield et al., 2021) and the post-processing of ensemble predictions (Hewson and Pillosu, 2021). Indeed, the early successes of these machine learning methods have led to the development of a 10-year roadmap for machine learning at ECMWF (Düben et al., 2021), with machine learning methods looking to be integrated into the operational workflow and machine learning demands considered in the procurement of HPC facilities; the ongoing development of novel computer architectures (e.g. GPU, IPU, FGPA) motivates utilizing algorithms and techniques which can efficiently take advantage of these new chips and gain significant performance returns. In this work we will demonstrate a new technique for the Verification of Earth-System ParametERisation (VESPER) based on a deep learning neural network regression model. This tool enables the accuracy of an updated water body parametrization to be rapidly and automatically assessed, and the added value that such an updated parametrization brings to be quantitatively evaluated.

This paper is organized as follows. In Section 2 we describe the construction of the VESPER tool - the raw input data, the processing steps and the construction of a neural network regressor. In Section 3.1 we then deploy VESPER to investigate and evaluate updated lake parametrisation fields. Discussion and concluding remarks are made in Sections **??** and **??** respectively.

2 Constructing VESPER

In order to rapidly assess accuracy of new surface physiography fields and if their use in the model increase predictability a neural network regression model (further in the text referred as VESPER) that can learn the mapping between a set of features x and targets y is constructed. In this case the features are the atmospheric and surface model fields (such as 2 metre temperature from ERA5 reanalysis) and the surface physiographic fields (such as orography and vegetation cover used to produce ERA5 reanalysis), see Table 1 for the full list of variables used. The target is the satellite land surface temperature (skin temperature from MODIS Aqua Day MYD11A1 v006 collection).VESPER can then make predictions about the skin temperature given a set of input variables (i.e. atmospheric and surface model fields, and surface physiographic fields). In turn, these predictions can then be compared against observations (i.e. satellite skin temperature) and VESPER's accuracy evaluated. By varying the number/type and values of the input features x to VESPER and observing how the accuracy of its predictions change, some conclusions on if and how features can increase predictability of an actual atmospheric model. Moreover, by isolating geographic regions where the predictions get worse with new/updated surface physiographic fields, areas where these fields might be erroneous or not informative enough can be identified. Due to the inherent stochasticity of training a neural network regression model it is also possible for different models to settle in different local minimums i.e. the network variance.

To understand the significance of this, every VESPER configuration was trained four times, each time with a different seed. The size of this variance is much smaller than the variance between different VESPER configurations

In this section we will now describe the data used for the features x and targets y in the neural network regression model, how various data types are joined together, and the details of VESPER's construction.

2.1 Features and targets

VESPER's input feature selection (see Table X) followed (i) permutation importance results for atmospheric and surface model fields - only fields with the highest importance were chosen; and (ii) expert choice for surface physiographic fields - as first attempt it was decided to test current methodology for lake related information, therefore fields that could be most affected by the presence of absence of water were selected, e.g. if lake had to be removed then some other surface had to appear (like bare ground, high or low vegetation, glacier or even ocean) and surface elevation had to change. Changes to the orographic fields will have important influences on temperature through e.g. wind, solar heating etc. Lake depth changes are similarly important, influencing how a lake freezes, thaws, mixes and its overall dynamical range. VESPER's target selection followed globally available criteria and the satellite land surface temperature is quite well observed globally and with high temporal pattern (daily or even several times a day depending on the location)

2.2 Data sources

TK:This section is new

There are three main sources of data. The first is selection of surface physiographic fields from ERA5 (Hersbach et al., 2020) and their updated versions (Choulga et al., 2019; Boussetta et al., 2021) (Muñoz Sabater et al., 2021) used as VESPER's

Main surface physiographic fields (19 fields)	Pressure: surface pressure (sp, Pa), mean sea level pressure (msl, Pa), Wind: 10 metre U wind component (10u, m/s), 10 metre V wind component (10v, m/s), Temperature: 2 metre temperature (2t, K), 2 metre dewpoint temperature (2d, K), skin temperature (skt, K), ice temperature layer 1 (the sea-ice temperature in layer 0-7 cm; istl1, K), ice temperature layer 2 (the sea-ice temperature in layer 7-28 cm; istl2, K), Surface albedo: forecast albedo (fal, 0-1), Snow: snow depth (sd, m of water equivalent)
Scenarie:	At oprette en server med bestemte regler som tillader folk at spille sammen. More Text more text More Text

features. The second is a selection of atmospheric and surface model fields from ERA5, also used as VESPER's features. The third is day-time land surface temperature measurements from the Moderate Resolution Imaging Spectroradiometer (MODIS) onboard the Aqua satellite (Center), used as VESPER's target variable.

2.2.1 Surface physiographic fields

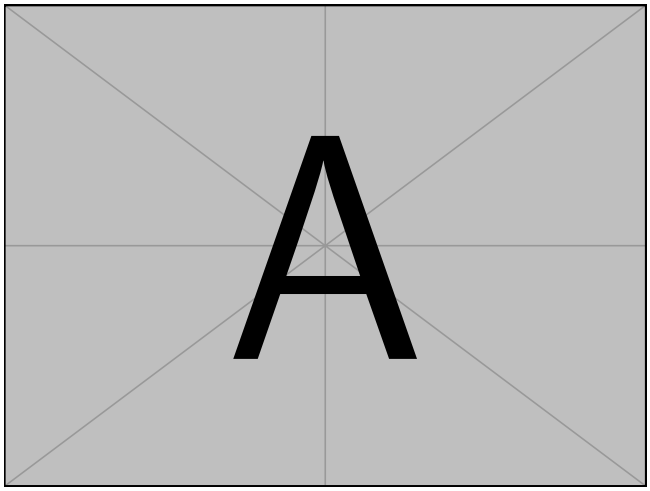
TK:This entire section is new

Surface physiographic fields have gridded information of the Earth's surface properties (e.g. land-use, vegetation type and distribution) and represent surface heterogeneity in the ECLand of ECMWF's Integrated Forecasting System (IFS). They are used to compute surface turbulent fluxes (of heat, moisture and momentum) and skin temperature over different surfaces (vegetation, bare soil, snow, interception and water) and then to calculate an area-weighted average for the grid-box to couple with the atmosphere. To trigger all different parametrization schemes the ECMWF model uses a sets of physiographic fields, that do not depend on initial condition or forecast step. Most fields are constant; surface albedo is specified for 12 months to describe the seasonal cycle. Dependent on the origin, initial data comes at different resolutions and different projections, and is then first converted to a regular latitude-longitude grid (EPSG:4326) at 1km at Equator resolution and secondly to a required grid and resolution. Surface physiographic fields used in this work consist of orographic, land, water, vegetation, soil, albedo fields and their difference between initial V15 and updated V20 field sets (and between expanded V15X and V20X field sets). See Tables 1 and 2 for the full list of surface physiographic fields and their input sources; for more details see IFS Documentation (2021).

As this work is focused on assessing quality of inland water information, main surface physiographic fields are lake cover (derived from land-sea mask) and lake mean depth (see Table 2). To generate V15 (and V15X) fractional lake cover the GlobCover2009 global map (Bontemps et al., 2011; Arino et al., 2012) is used. This map has a resolution of 300m, corresponds for the year 2009 and covers latitudes 85°N-60°S; corrections outside these latitudes for the polar regions are included separately. In the Arctic no land is assumed, in the Antarctic data from the high-resolution Radarsat Antarctic Mapping Project digital elevation model version 2 (RAMP2; Liu et al., 2015) is used. To generate V20 (and V20X) fractional lake cover more resent higher resolution datasets and updated methods have been used (Choulga et al., 2019). Main data source is the Joint Research

Centre (JRC) the Global Surface Water Explorer (GSWE) dataset (Pekel et al., 2016). GSWE is a 30m resolution dataset from Landsat 5,7 and 8, providing information on the spatial and temporal variability of surface water on the Earth since March 1984; here only permanent water type was used for lake cover generation as it was more accurate inland water distribution on the annual basis (Choulga et al., 2019). Differences between V20 and V15 lake cover fields (see Figure 1) are consistent with the latest global and regional information: (i) increase of lake fraction in V20 compared to V15 over northern latitudes is due to permafrost melt leading to a new thermokarst lake emergence, and due to higher resolution input source and its better satellite image recognition methodologies; (ii) reduction of lake fraction in V20 compared to V15 can be explained with several reasons, like anthropogenic land use change (e.g. Aral Sea, which lies across the border between Uzbekistan and Kazakhstan, has been shrinking at an accelerated rate since the 1960s and started to stabilise in 2014 with an area of 7660km2, 9 times smaller than its size in 1960. GlobCover2009 describes the Aral Sea in 1998, when it was still "only" two times smaller than its 1960 extent, whereas GSWE provides a more up to date map.), use of only permanent water (e.g. Australia, where GlobCover2009 over-represents inland water, as most of these lakes are highly ephemeral, e.g. the endorheic Kati Thanda–Lake Eyre fills only a few times per century. The GSWE updates to this region therefore include only generally permanent water, removing all seasonal and rare ephemeral water.), and change in the ocean and inland water separation algorithm (e.g. north-east of Russia). To generate V15 (and V15X) lake mean depth (see Figure 2a) the Global Lake DataBase version 1 (GLDBv1; Kourzeneva et al., 2012) is used. GLDBv1 has a resolution of 1km and is based on 13000 lakes with in situ lake depth information; outside this dataset all missing data grid-cells (i.e. over ocean and land) have 25 meter value; field aggregation to a coarser resolution is done by averaging. Overestimation of lake depth in summer season can result in strong cold biases and in winter season – lack of ice formation. To generate V20 (and V20X) lake mean depth (see Figure 2b) an updated version GLDBv3 (Choulga et al., 2014) is used. GLDBv3 has the same resolution of 1km, but is based on an increased number of lakes with in situ lake depth information by $\sim$1500 (in addition has bathymetry information over all Finnish navigable lakes), it introduces distinction between freshwater and saline lakes (this information is currently not used by FLake), and suggests the method to assess the depth for lakes without in situ observations using geological and climate type information; field aggregation to a coarser resolution is done by computing the most occurring value. Verification of GLDBv1 and GLDBv3 lake depths against 353 Finnish lake measurements shows that GLDBv3 exhibits a 52 % bias reduction in mean lake depth values compared to GLDBv1 (Choulga et al., 2019). For a further details on lake distribution and depth, the representation of lakes by ECMWF in general see Choulga et al. (2019) and Boussetta et al. (2021).

To expand V15 and V20 lake description (to V15X and V20X respectively, see Table 2) their salinity and time variability information was generated. Even though static permanent water fits better to describe inland water distribution on average all year round, some areas (in Tropics especially) would really benefit from having monthly varying information as they have a very strong seasonal cycle, when size, shape and depth of a lake changes over the course of the year, leading to a significant changes in modelling the lake temperature response. Similarly, saline lakes behave very differently to fresh water lakes since increased salt concentrations affect the density, specific heat capacity, thermal conductivity, and turbidity, as well as evaporation rates, ice formation and ultimately the surface temperature. These two properties of time variability and salinity are often related; it is common for saline lakes to fill and dry out over the course of the season, which naturally also affects the



(a)

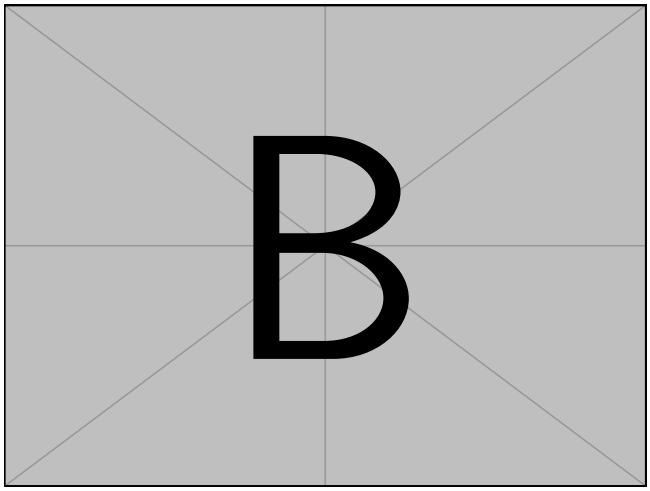
Figure 1. *TK: This figure is a placeholder for Figure 1 in the word doc*

relative saline concentration of the lake itself. To create a monthly varying lake cover first 12 monthly fractional land-sea masks based on JRC Monthly Water History v1.3 maps for 2010-2020 were created. Since the annual lake maps were created taking into account a lot of additional sources the extra condition on the monthly maps that the monthly water is equal or greater than permanent water distribution from fractional land-sea mask is enforced. To create an inland salt lake cover GLDBv3 salt lake list was used. First, in order to identify separate lakes on 1km resolution lake cover (here max lake distribution based on monthly varying lake cover is used to ease further application), small sub-grid lakes and large lake coasts are masked, i.e. grid-cells that have water fraction less than 0.25. Next, number of connected grid-cells in each lake (i.e. connected with sides only) is computed. Then only lakes that have 100 and more connected grid-cells are vectorised, as at ERA5 resolution of ∼31km the grid-cells are quite large and can include a mixture of freshwater and saline lakes. Finally, saline lake vectors are selected by filtering vectors which have no saline lake point from GLDBv3 located – in total 147 large salt lake vectors, which were further used to filter non-saline lakes at 1km resolution lake cover, finally aggregated to 31km resolution. In the future it is planned to revisit this field and extend the list to include additional data.

2.2.2 ERA5

TK:This section has been updated

Climate reanalyses combine observations and modelling to provide calculated values of a range of climactic variables over time. ERA5 is the fifth generation reanalysis from ECMWF. It is produced via 4D-Var data assimilation of the atmospheric Integrated Forecast system cycle 41R2, coupled to a land-surface model (ECLand, Boussetta et al., 2021), which includes lake parametrization by the Fresh water Lake model Flake (Mironov, 2008) and an ocean wave model (WAM). The resulting data product provides hourly values of climatic variables across the atmosphere, land and ocean at a resolution of approximately



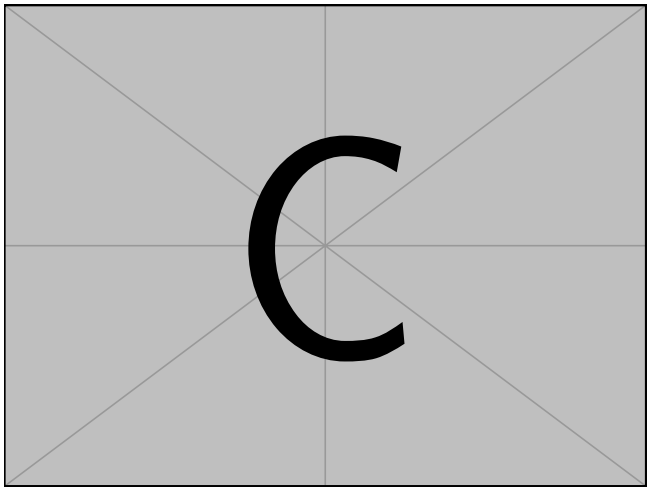
(a)

Figure 2. *TK: This figure is a placeholder for Figure 2 in the word doc*

Main surface physiographic fields (19 fields)	Pressure: surface pressure (sp, Pa), mean sea level pressure (msl, Pa), Wind: 10 metre U wind component (10u, m/s), 10 metre V wind component (10v, m/s), Temperature: 2 metre temperature (2t, K), 2 metre dewpoint temperature (2d, K), skin temperature (skt, K), ice temperature layer 1 (the sea-ice temperature in layer 0-7 cm; istl1, K), ice temperature layer 2 (the sea-ice temperature in layer 7-28 cm; istl2, K), Surface albedo: forecast albedo (fal, 0-1), Snow: snow depth (sd, m of water equivalent)
Scenarie:	At oprette en server med bestemte regler som tillader folk at spille sammen. More Text more text More Text

Table 1. *TK: This table is a placeholder for Table 2 in the word doc*

31km with 137 vertical sigma levels, up to a height of 80km. Additionally, ERA5 provides associated uncertainties of the variables at a reduced 63km resolution via a 10-member Ensemble of Data Assimilations (EDA). In this work ERA5 hourly surface fields at $\sim$ 31km resolution on a reduced Gaussian grid are used. Gaussian grid's spacing between latitude lines is not regular, but lines are symmetrical along the Equator; the number of points along each latitude line defines longitude lines, which start at longitude 0 and are equally spaced along the latitude line. In a reduced gaussian grid, the number of points on each latitude line is chosen so that the local east-west grid length remains approximately constant for all latitudes (here Gaussian grid is N320, where N is the number of latitude lines between a Pole and the Equator). Main field used from ERA5 is skin temperature (i.e. temperature of the uppermost surface layer, which has no heat capacity and instantaneously responds to changes in surface fluxes) that forms the interface between the soil and the atmosphere. Skin temperature is a theoretical temperature computed by linearizing the surface energy balance equation for each surface type separately, and its feedback on



(a)

Figure 3. *TK: This figure is a placeholder for Figure 3 in the word doc*

net radiation and ground heat flux is included; for more information see IFS Documentation (2021). ERA5 skin temperature verification against MODIS LST ensemble (i.e. all four MODIS observations were used, namely Aqua Day and Night, Terra Day and Night) over 2003-2018 period showed good correlation between two datasets; errors between ERA5 and MODIS LST ensemble are quite small, i.e. spatially and temporally averaged bias is 1.64K, RMSE is 3.96K, Pearson correlation coefficient is 0.94, and anomaly correlation coefficient is 0.75 (Muñoz-Sabater et al., 2021). ERA5 skin temperature verification against the Satellite Application Facility on Land Surface Analysis (LSA-SAF) product over Iberian Peninsula showed a general underestimation of daytime LST and slightly overestimation at night-time, relating the large daytime cold bias with vegetation cover differences between ERA5 surface physiography fields and the European Space Agency's Climate Change Initiative (ESA-CCI) Land Cover dataset; use of ESA-CCI low and high vegetation cover instead of ERA5 ones led to a complete reduction of the large maximum temperature bias during summer (Johannsen et al., 2019). ERA5 data is obtained via the Copernicus Climate Data Store (CDS, Muñoz-Sabater, 2019).

2.2.3 Aqua-MODIS

TK:This section has been updated

Aqua (Parkinson, 2003) is a NASA satellite mission which makes up part of the Earth Observing System (EOS). Operating at an altitude of 700km, with orbital period of 99 minutes, its orbital trajectory passes south to north with an equatorial-crossing times in general of 1.30pm. This post-meridian crossing time has led to it sometimes being denoted as EOS PM. Launched in 2002 with an initial expected mission duration of 6 years, Aqua has far exceeded its initial brief and continues to transmit information from 4 of the 6 observation instruments on board. Here we use information only from MODIS instrument. MODIS can take surface temperature measurements at a spatial resolution of 1km (the exact grid size is 0.928km by 0.928km), operating

in the wavelength ranges of between ∼3.7-4.5μm and ∼10.9-12.3μm. In addition to surface temperature measurements that were used in this work, MODIS can take observations of cloud properties, water vapour, ozone etc. Here MYD11A1 v006 (Wan et al., 2022) collection that provides daily Land Surface Temperature (LST) measurements at a spatial resolution of 1km on a sinusoidal projection grid SR-ORG:6974 (takes a spherical projection but a WGS84 datum ellipsoid) is exercised. Daily global LST data is generated by first applying a split-window LST algorithm (Wan and Dozier, 1996) on all nominal (i.e. 1km at nadir) resolution swath (scene) with a nominal coverage of 5 minutes of MODIS scans along the track acquired in daytime, and secondly by mapping results onto integerized sinusoidal projection; for more details see Figure 3 and Wan (2019). Validation of this product was carried out using temperature-based method over different land cover types (e.g. grasslands, croplands, shrublands, woody areas, etc.) in several regions around the globe (i.e. United States, Portugal, Namibia, and China) at different atmospheric and/or surface conditions; the best accuracy is achieved over United States sites with RMSE lower than 1.3K (Duan et al., 2019). At large view angles and in semi-arid regions product may have slightly higher errors due to rather uncertain classification-based surface emissivities and heavy dust aerosols effects. In addition, MODIS Cloud Mask struggles to eliminate all cloud and/or heavy aerosols contaminated grid-cells from the clear-sky ones (LST errors in such grid-cells can be 4-11K and larger). Validation of this product over five bare ground sites in north Africa (in total 12 radiosonde-based datasets validated) showed that mean LST error was within ±0.6K (with exception for one dataset, where mean LST error was 0.8K) and standard deviation of LST errors were less than 0.5K (Duan et al., 2019). In this work to reduce the amount of daily data over multiple years to store and manipulate, prior use LST data is (i) filtered to contain only cloud free data, and (ii) averaged to a 4km at the Equator resolution on a regular latitude-longitude grid, EPSG 4326 (note that only grid cells which have 8 or more valid observations at 1km resolution are averaged over, otherwise they are classified as missing data).

2.3 Joining the data

To join selected ERA5 global fields on a reduced Gaussian grid at 31km resolution (information in UTC, 24 hourly maps per day) with Aqua-MODIS global LST data on a regular latitude-longitude grid at 4km resolution (information in local solar time, 1 map per day), both datasets need to be at the same time space. First it is necessary to determine the absolute time (i.e. UTC) at which the MODIS observations were taken. Since in general all Aqua observations are taken at 1.30pm local solar time, it can be related to a UTC via observation longitude, following Eq. (1):

$$\text{UTC} = \text{Local solar time} - \frac{\text{longitude}}{15}\ , \tag{1}$$

where longitude is in degrees, and UTC is rounded to the nearest hour. This conversion is inexact since there is an additional correction as a function of the latitude, yet recommended by the official MODIS Products User's Guide (Wan, 2019); given the short orbital period of Aqua these additional higher order corrections are expected to be typically small and for our purposes can be neglected. Also, assumption that all Aqua observations are taken at 1.30pm local solar time was checked (see Figure 4). The annually averaged mean time difference at 31km resolution (i.e. daily differences between local solar time of observations and 1.30pm at 1km resolution were first aggregated to 31km resolution using averaging, and then aggregated in time over a year) is 0.16 hours or 10 min, with mean absolute error (MAE) being 0.46 hours or 28 min and root mean square error (RMSE)

being 0.61 hours or 37 min (current values correspond 70N-70S region year 2019, but confirmed to be approximately identical for each year of 2016-2019 period). Since temporal resolution of ERA5 data is hour, proposed assumption has sufficient accuracy. Over the poles (i.e. 90-70°N and 70-90°S) satellite sweeps overlap significantly and in general conversion becomes less accurate (daily time differences can reach more than ± 3.5 hours), so these areas were not included in the analysis. Once Aqua-MODIS time of observation is converted to UTC, Aqua-MODIS data at 4km resolution is matched in time and space to ERA5 information in a following way: (i) take a single Aqua-MODIS LST observation at a particular point on the MODIS grid, (ii) select ERA5 global hourly map matching Aqua-MODIS LST observation time in UTC, (iii) find the nearest point on the ERA5 grid to that MODIS grid point, (iv) repeat previous steps for every Aqua-MODIS observation, (v) group matched data pairs by the ERA5 grid points, averaging over all the Aqua-MODIS observations that are associated with each ERA5 point. At the end selected ERA5 fields are mapped to a single Aqua-MODIS time of observation and Aqua-MODIS LST data is mapped (i.e. multiple Aqua-MODIS observations could be averaged over, see Figure 5a) to a reduced Gaussian grid at 31km resolution; averaged Aqua-MODIS observations are considered as ground truth (i.e. targets y) that VESPER is trying to predict. To better understand VESPER's grid-cell results at 31km resolution additional information was computed from Aqua-MODIS, namely (i) total number of valid observations per month and year (see Figure 5a), and (ii) average LST error based on Aqua-MODIA quality assessment (i.e. quality flag, see Figure 5b). Based on this additional information it can be concluded that areas with sparse number of observations in general have more uncertain LST values; exceptions are Alaska in United States and Anadyrsky District in Russia (area 30° east and west from 180°E around 70-60°N), deserts of Australia and Kalahari desert in Namibia, Botswana and South Africa, where majority of vast number of observations have only good or average quality.

Second step in joining ERA5 and Aqua-MODIS LST data uses a GPU-accelerated k-nearest neighbours algorithm (RAPIDS, v22.04.00), where "nearness" on the sphere between two points is measured via the Haversine metric i.e. the geodesic distance H, following Eq. (2):

$$H = 2\arcsin(d) \tag{2}$$

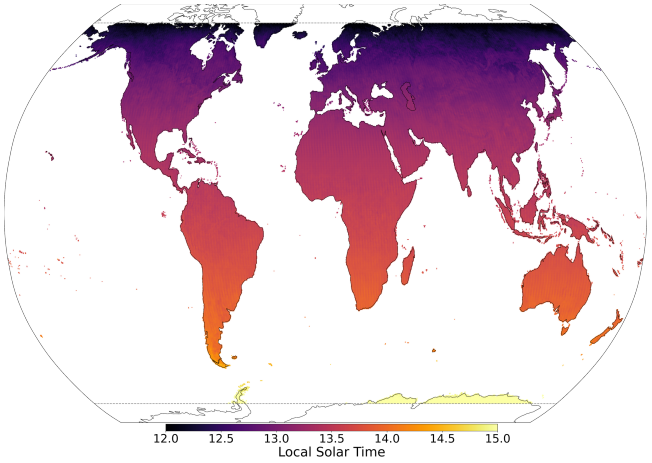
where

$$d = \sqrt{\sin^2\left(\frac{\delta\theta}{2}\right) + \cos\theta_1 \cos\theta_2 \sin^2\left(\frac{\delta\phi}{2}\right)} \tag{3}$$

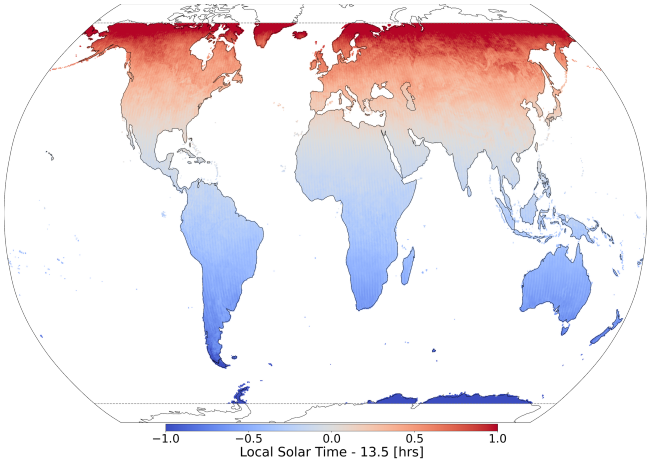
for two points with coordinate latitudes $\theta_{1,2}$, longitudes $\phi_{1,2}$ and $\delta\theta = \theta_2 - \theta_1$ and $\delta\phi = \phi_2 - \phi_1$.

2.4 Constructing a regression model

VESPER is trained to learn the mapping between features x and targets y (i.e. mapping ERA5 to MODIS), a regression problem. For this purpose a fully-connected neural network architecture (also known as a multi-layer perceptron), implemented in Tensorflow (Abadi et al., 2015) was used. Whilst more advanced architectures are available, for the purposes of this work the model is sufficient enough, which exhibits generally fast and dependable convergence. The networks built have differing number of nodes in the input layer, depending on the number of predictors (see table below). For all networks constructed



(a)



(b)

Figure 4. Average *(a)* Local solar time of MODIS Aqua day, *(b)* Error relative to the assumed local solar time of 13.30 for the year 2019 at a 31km resolution. The errors are generally sub-hr and grow at greater latitudes. We exclude data with latitudes $|\theta| < 70^{\circ}$ and take 13.30 as a constant local solar time.

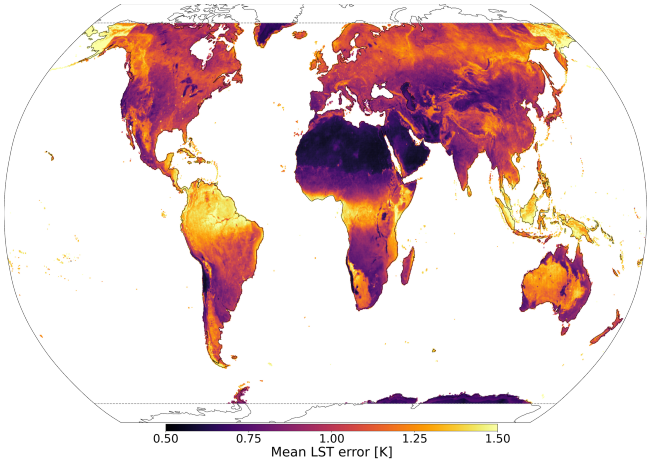


Figure 5. Average error in the MODIS LST measurement at a 31km resolution. The raw MODIS data at a 1km resolution provides categorical LST errors with bins ≤ 1K, $1-2$K $2-3$K and >3K. When averaging to 31km resolution we compute a weighted average over the 1km grid cells, where we take the median bin value, and 5K for the >3K bin.

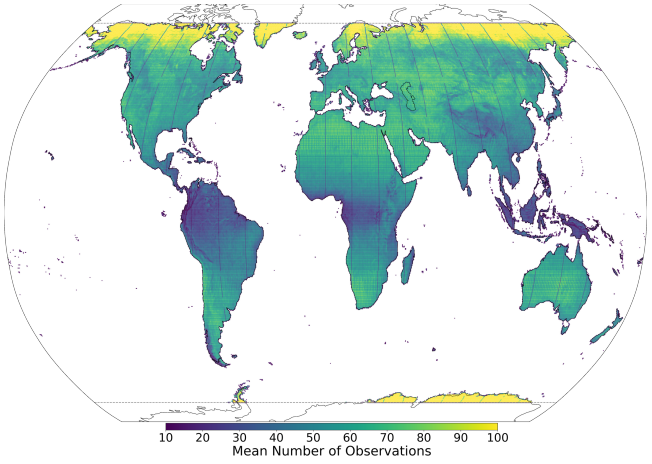


Figure 6. Mean daily number of MODIS observations mapped to each ERA5 data point for 2019. The swath of the Aqua satellite is clearly visible, with more observations at more extreme latitudes as Aqua follows a polar orbit, south to north. In addition to the expected increased sparsity of observations at the equator, there are also notably fewer observations in regions of greater orography such as the Himalayas, the Andes and the Rocky Mountains as well as the Siberian Tundra, due to increased cloud cover.

we use 4 hidden layers and a layer width is half that of the input layer width. ADAM (Kingma and Ba, 2014) is used as an optimisation scheme, learning rate is set to 3×10-4, and default values for the exponential decay rate for the 1st and 2nd moment estimates are set to of 0.90 and 0.999 respectively. The network is not trained for a fixed number of epochs, but instead trained until the validation error reaches a minimum. Techniques for maximising the performance of a network via hyperparameter optimisation are now well established (Bischl et al., 2021; Yu and Zhu, 2020). However, for the purposes of this work no attempt to tune hyperparameters was made, just some reasonable default values were applied which were assumed to be "good enough". Some exploration of different hyperparameter configuration was undertaken, but for this data the prediction accuracy is mostly independent of the hyperparameter configuration, subject to standard and reasonable hyperparameter choices. Whilst a more advanced automatic hyperparameter optimization method may have enabled slightly higher performance of VESPER, ultimate purpose is not to generate the most absolutely accurate prediction possible, but instead to have two predictive models which can be compared. In the result section below it will be shown that the variation in performance due to input feature modifications is far greater than the variation due to the hyperparameter choices. Here VESPER was trained on selected atmospheric and surface model fields from ERA5 (see Table 1) and Aqua-MODIS LST for 2016 and certain static version of the surface physiographic fields (see Table 2), once VESPER was fully trained it was used to predict LST over the whole globe for 2019. Finally VESPER's predictions can be compared to initial ERA5 skin temperatures and actual Aqua-MODIS LST for 2019 (see Figure 6; VESPER trained using V15 surface physiographic fields is called VESPERV15, for full explanation see Table 3). Figure 6 shows mean absolute errors (MAE) of differences between VESPERV15 LST predictions and Aqua-MODIS LST (see Figure 6a), and between ERA5 skin temperature and Aqua-MODIS LST (see Figure 6b); based on that a conclusion can be drawn that VESPERV15 was able to learn corrections to ERA5, especially in the Himalayas and sub-Saharan Africa as well as Australia and the Amazon basin, leading to the globally averaged MAE reduction for predicted LST (MAEERA5=3.9K comparing to MAEVESPERV15=3.0K).

All VESPER versions are trained with ERA5 fields for 2016 and with main surface physiographic fields V15. Then depending on the version some or all additional surface physiographic fields (see Table 1) are added. Note that all non-lake related climate fields such as vegetation cover or orography were updated in V20 field set comparing to V15 only in relation to the changing lake fields (i.e. if fraction of lake in the grid cell increased then other fractions like vegetation or bare ground should have increased accordingly).

As focus of this study are lake related fields, and lakes occupy only 1.8% of the Earth's surface and are distributed very heterogeneously (Choulga et al., 2014), analysis of the results here was restricted to areas where there have been significant changes in the surface physiographic fields, i.e. a change in any of the surface field when going from V15 to V20 (and to V15X or V20X) of $\geq$ 1% ($\geq$ 0.1 for fractional fields); if lake or vegetation cover changed from 0.1 in V15 to 0.3 in V20 field set this change is classified as significant. Choice of $\geq$ 10 % cut-off was adopted as it proofed to be a good trade off between having a sufficient number of grid points to inspect and the strength of the effect of changing the input field. As the cut-off % increases less points are selected even though with more severe changes, whereas when the cut-off % decreases more points are selected but it becomes more difficult to disentangle the change in the prediction accuracy from VESPER's training noise. Alternative cut-off % were briefly explored, but conclusions of the results remained broadly unchanged. All

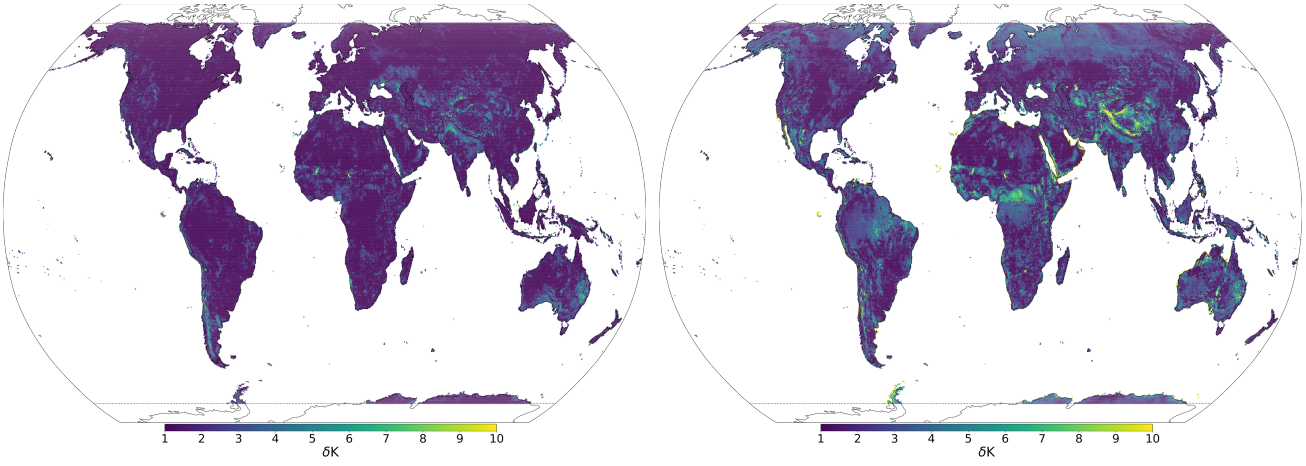
grid-cells selected for the analysis can be classified according to how the surface fields are updated when going from V15 to V20 (note that categories represent a systematic and consistent update across multiple related fields, and do not include any restrictions on other surface fields apart the ones mentioned): – Lake Updates. The change in the lake cover cl and lake depth dl are significant, but the changes in ocean and glacier glm fractions are not. This corresponds to grid-cells where lakes have been added or removed. Lake-Ground Updates is a sub-category where additional constraint that the change in the high/low vegetation fractions cvh/cvl are not significant is in place. This then corresponds to the exchange of lakes for bare ground, or vice versa. – Vegetation Updates. The change in the high vegetation fraction cvh is significant, but the change in lake cover cl is not significant. This corresponds to grid-cells where large features like forests and woodlands have been updated, exchanged for bare ground or low vegetation. – Glacier Updates. The change in the glacier cover glm is significant. This corresponds to any areas where the fraction of glacier ice has been updated. All four VESPER versions were trained on 2016 data over the entire globe, and then used to predict LST for 2019. To compute the training noise (i.e. the network variance – changes in the LST predictability due to model training) every VESPER's version was trained four times and each time with a different seed. To assess the changes of LST predictability due to the use of the updated surface physiographic fields instead of V15 field set (default) a simple metric VM is computed, following Eq. (3):

$$1 + 1 \tag{4}$$

where M represents one of the field set versions V20, V20X or V15X, and MAE is computed over the whole prediction period of 2019; e.g. V20 describes the difference between MAE of differences between VESPERV20 LST predictions and Aqua-MODIS LST and between MAE of differences between VESPERV15 LST predictions and Aqua-MODIS LST, negative V20 indicates that VESPERV20 LST prediction is more accurate and vice versa.

Main surface physiographic fields (19 fields)	Pressure: surface pressure (sp, Pa), mean sea level pressure (msl, Pa), Wind: 10 metre U wind component (10u, m/s), 10 metre V wind component (10v, m/s), Temperature: 2 metre temperature (2t, K), 2 metre dewpoint temperature (2d, K), skin temperature (skt, K), ice temperature layer 1 (the sea-ice temperature in layer 0-7 cm; istl1, K), ice temperature layer 2 (the sea-ice temperature in layer 7-28 cm; istl2, K), Surface albedo: forecast albedo (fal, 0-1), Snow: snow depth (sd, m of water equivalent)
Scenarie:	At oprette en server med bestemte regler som tillader folk at spille sammen. More Text more text More Text

Table 2. *TK: This table is a placeholder for Table 3 in the word doc*



(a) (b)

Figure 7. Prediction error relative to MODIS Aqua observations in the land surface temperature (δK) for 2019, averaged over the year, for *(a)* Trained Neural Network and *(b)* ERA5. It can be seen that the network generally outperforms the ERA5 predictions, which generally struggles in regions with complex surface fields such as the Himalayas (lots of orography) sub-Saharan Africa (lots of vegetation) and the Amazon Basin (lots of water + vegetation). In contrast the network demonstrates generally good performance, with some drop off in the Himalayas and the eastern cost of Australia, but still outperforming ERA5.

3 Results

3.1 Evaluation of updated lake fields

To understand if there is a way to automatically and rapidly assess accuracy of updated and/or new surface physiography fields and if their use in the atmospheric model increase predictability a neural network regression model VESPER was trained over 2016 data and used to predict LST for 2019. VESPER's training noise was confirmed to be much smaller (estimated noise bias over all grid-cells for 2019 at 31km resolution is 0.04K) than differences in LST predictions by different VESPER configurations (degrees K), so all changes in LST predictability can be attributed to the changes in surface physiographic fields. As a first attempt lake related information is assessed, namely lake cover (and land-sea mask and glacier cover as they are used for lake cover generation) and lake mean depth, that were created from scratch using new up-to-date high-resolution input datasets (see Table 2) for the V20 (and V20X) field set; other surface physiographic fields (see Table 1) were regenerated from the same input sources as in initial V15 field set but taking into account that lake related fields were changed. In cases when existing in V15 lake cover water was removed in V20, its place could fill high or low vegetation, glacier or bare ground. To analyse results grid-cells were divided into four categories, and each was investigated in detail (see Table 4 for the results aggregated over the whole globe).

Main surface physiographic fields (19 fields)	Pressure: surface pressure (sp, Pa), mean sea level pressure (msl, Pa), Wind: 10 metre U wind component (10u, m/s), 10 metre V wind component (10v, m/s), Temperature: 2 metre temperature (2t, K), 2 metre dewpoint temperature (2d, K), skin temperature (skt, K), ice temperature layer 1 (the sea-ice temperature in layer 0-7 cm; istl1, K), ice temperature layer 2 (the sea-ice temperature in layer 7-28 cm; istl2, K), Surface albedo: forecast albedo (fal, 0-1), Snow: snow depth (sd, m of water equivalent)
Scenarie:	At oprette en server med bestemte regler som tillader folk at spille sammen. More Text more text More Text

Table 3. *TK: This table is a placeholder for Table 4 in the word doc*

3.1.1 Category: Lake updates

Lake Updates category shows significant improvements in LST predictability if using V20 field set instead of V15 (see Table 4 and Figure 7) – prediction accuracy increased globally (over 1631 grid-cells) on average by 0.45K (training noise in VESPERV15 XX, and in VESPERV20 XX), most notably in Australia and the Aral sea (two major regions where ephemeral lakes were removed and inland water distribution made up-to-date respectively, discussed in Surface physiographic fields section).

In addition to the areas with a notable improvement in the prediction accuracy, there are some noteworthy regions where the predictions got worse (see red points in Figure 7) suggesting inaccuracies or lack of information in the updated surface physiographic fields. A few of the most noteworthy grid-cells (see red points highlighted with green circles in Figure 7) are: – Tanzania. There are two grid-cells at the centre and northern edge of Lake Natron, in Tanzania (lies to the south-east of Lake Victoria) where VESPERV20 LST predictions are less accurate than VESPERV15 (training noise in VESPERV15 XX and XX, and in VESPERV20 XX and XX, for the central and northern edge of the lake respectively). For the central point (see Figure 8a) the lake fraction was increased from 0.04 in V15 to 0.39 in V20 field set; for the northern edge point lake fraction was also increased in V20 comparing to V15 field set along with a small decrease ($\sim$0.1) in the low vegetation fraction. However, Lake Natron is a highly saline lake that often dries out, with high temperatures, high levels of evaporation and irregular rainfall. It is a highly complex and variable regime that is not well described by simply increasing fraction of permanent fresh water, and indeed results suggest that with current lake parametrization scheme it may be beneficial to keep the lake fraction low or introduce extra descriptor, e.g. salinity. – Australia. Grid-cell contains Lake Blanche, in South Australia (see Figure 8b). Lake fraction was completely removed from 0.44 in V15 to 0.00 in V20 field set, along with lake depth reduction from 5.5m to 1.0m, and low vegetation fraction increase from 0.53 to 0.97. Whilst the removal of ephemeral water is generally accurate for Australia, for this grid-cell it lowers the predictability of VESPERV20 comparing to VESPERV15 (training noise in VESPERV15 XX, and in VESPERV20 XX). Lake Blanche is a salt lake that lies below sea level within a wetlands system (so will retain some surface water or soil moisture which will influence the temperature response), fairly devoid of any obvious vegetation. V20 field set describes grid-cell as above sea level (same in V15 orography fields) completely dry region covered with short grass (i.e. low vegetation), which is not an accurate description resulting in worse predictions. – Salt Lake City, North America. Grid-cell lies within the Great Salt Lake Desert, just to the west of the Great Salt Lake, Utah, US. Lake fraction was completely removed from GTR 0.50 in V15 to 0.00 in V20 field set, grid-cell is fully covered with bare ground in V20 field set (training noise in VESPERV15 XX, and in VESPERV20 XX). Whilst this area primarily is bare ground, satellite imagery also suggests the presence of a presumably highly saline lake (see Figure 6c); in addition area has a large degree of orography and high elevation ($\sim$1300m) which probably further complicates the surface temperature response. A more accurate description that accounts for the seasonality of the surface water and the salinity is necessary here. – Afghanistan. Grid-cell lies in the south-west of Afghanistan, close to the border with Iran. Lake fraction was completely removed from 0.11 in V15 to 0.00 in V20 field set (training noise in VESPERV15 XX, and in VESPERV20 XX). However, this area in fact has an extensive network of mountain tributaries which feed an ephemeral lake (e.g. see Figure 6d). Most likely there should be some inland water within a year, especially during the rainy season, so complete water removal might be an overcorrection.

– Northern India. Grid-cell lies in the state of Gujarat, India, close to the border with Pakistan. Lake fraction was increased from 0.59 in V15 to 0.71 in V20 field set, along with the lake depth increase from 2.58m to 3.76m (training noise in VESPERV15 XX, and in VESPERV20 XX). However, this point lies on a river delta within the Great Raan of Kutch, a large area of salt marshes (see Figure 6e), known for having highly seasonal rainfall, with frequent flooding during the monsoon season and a long dry season. The surface itself also undulates with areas of higher sandy ground known as medaks, with greater levels of vegetation. It is evidently a complex and highly time variable area and additional static fraction of fresh water provided via V20

field set is not enough. – Egypt. Grid-cell contains a section of the Toshka Lakes, lies to the west of the River Nile in the south of Egypt. Lake fraction was completely removed from 0.36 in V15 to 0.00 in V20 field set, along with the lake depth decrease from 25m to 6m (training noise in VESPERV15 XX, and in VESPERV20 XX). However, whilst this is a very dry region, it contains part of Toshka Lakes, a collection of endorheic waterbodies newly formed (and growing) due to overflow from Lake Nasser (see Figure 6f). These lakes are known to be highly time variable, with a periodic seasonality on top of the general increasing lake sizes, and the formation of surrounding wetlands. These lakes rapidly fill and dry out; during the training and validation years of VESPER (along with the decade before) lakes were mostly dry, whereas during the testing year they were filled. It is worth mentioning a vast area north-west of Canada where VESPERV20 slightly underperforms VESPERV15 (V20 averaged over the area is +0.02K; training noise in VESPERV15 XX, and in VESPERV20 XX), but it is hard to draw any definitive conclusions – whilst LST predictions for some grid-cells become more accurate, for some – less accurate. For these high latitude regions there is a large uncertainty for lake location (thermokarst lakes emerge due to permafrost thawing, have small area and can quickly disappear) and lake depth which influence water freezing/thawing dates and ice duration. In addition, observations in these regions are more uncertain and have greater errors due to increased cloud cover. Another interesting region is the north-eastern edge of the Caspian Sea, where VESPERV20 underperforms VESPERV15 (V20 averaged over four grid-cells is +0.65K; training noise in VESPERV15 XX, and in VESPERV20 XX). Location is the Astrakhan Nature Reserve (an extensive wetland). In general, lake fraction was reduced and vegetation fraction increased accordingly in V20 comparing to V15 field set. However, it seems that permanent fresh water fraction is not enough to accurately describe the region and additional monthly varying water cover and/or wetland cover should be introduced.

Lake-Ground Updates sub-category, which restricts analysis to only points with no significant change in the vegetation, allows to more clearly see the effect of adding/removing water on/from bare ground. Sub-category shows even bigger improvements in LST predictability if using V20 field set instead of V15 (see Table 4) – prediction accuracy increased globally (over 546 grid-cells) on average by 1.12K (training noise in VESPERV15 XX, and in VESPERV20 XX). This indicates that whilst the updated lake fields are globally accurate and informative, providing on average over the globe, over a year, more than an extra Kelvin of predictive performance, the updates to the vegetation fields tamper this performance gain, indicating potential problem with the vegetation fields

3.1.2 Category: Vegetation Updates

Vegetation Updates category, which restricts analysis to only points with significant change to high vegetation cover when it's substituted with low vegetation or bare ground and vice versa (no significant change in the lake cover), allows to narrow down the drop of LST predictability issue in VESPERV20 comparing to VESPERV15 due to vegetation change (see Table 4) – prediction accuracy decreased globally (over 58 grid-cells only) on average by 0.49K (training noise in VESPERV15 XX, and in VESPERV20 XX). For all grid-cells in this category apart from one, the high vegetation fraction was decreased (often quite drastically to zero), specifying that there should just be bare ground. Thorough inspection of these areas with satellite imagery revealed that they should in fact be covered with high vegetation (see Figure 9 for an example) and that update in V20 high vegetation cover was erroneous for these grid-cells. It was noted that the strength of drop in LST predictability in VESPERV20

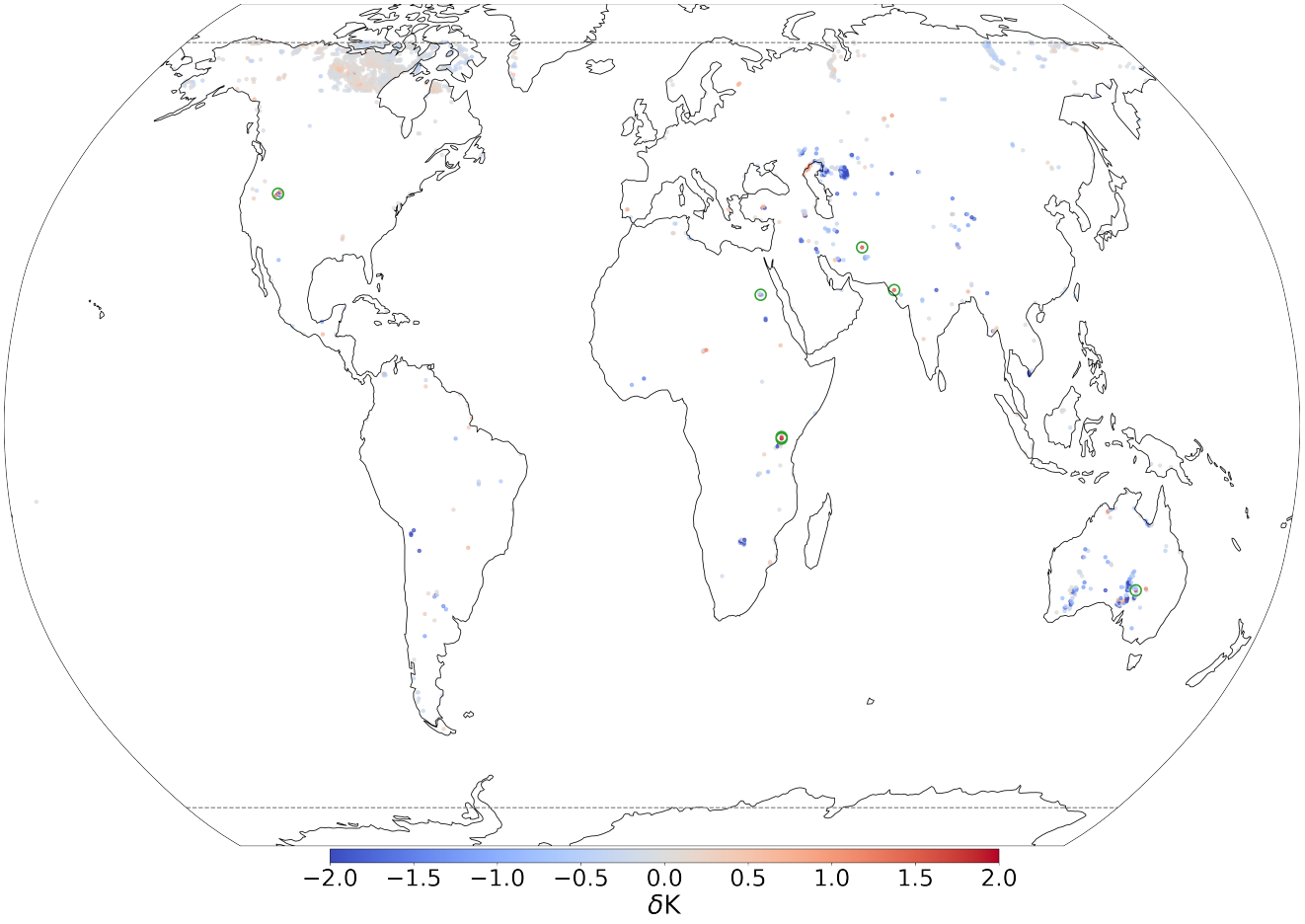
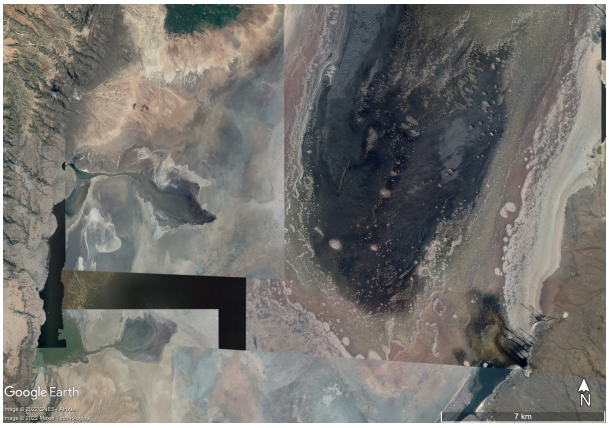
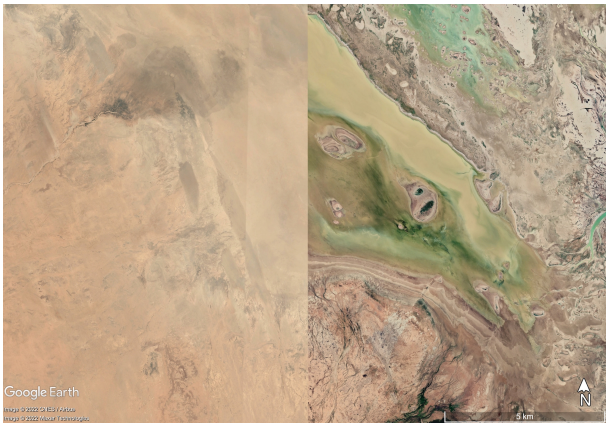


Figure 8. Mean δ for the V20 model relative to the V15 model across the globe for grid points where all the lake fields have changed significantly ("Lake Update" category). Generally the updated V20 fields enable the model to make more accurate predictions, for example in the Aral sea and Australia, indicating that these updated fields are informative and accurate. In contrast, there are some regions where the predictions get worse, for example at higher latitudes which is likely due to these being regions where lakes have more complex, time variable behaviour (e.g. freezing/thawing) and MODIS satellite data is sparse e.g. due to clouds. 7 points (two are overlaid in sub-Saharan Africa) where the V20 prediction gets notably worse than V15 are highlighted with green circles and discussed in the text.

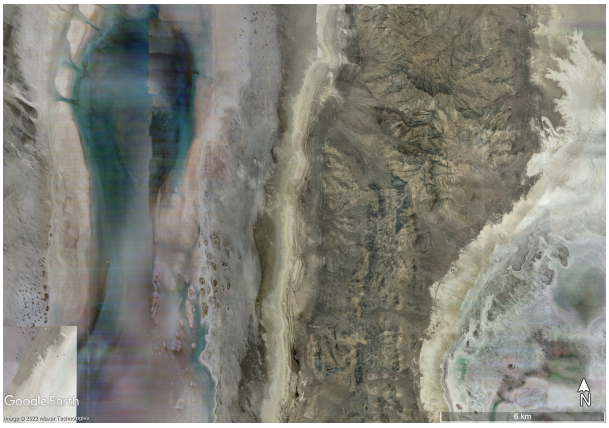
Differences V20 between mean absolute error (MAE) of [VESPERV20 LST predictions minus Aqua-MODIS LST] and MAE of [VESPERV15 LST predictions minus Aqua-MODIS LST] for 2019 at 31km resolution for 'Lake Updates' category (i.e. where lake cover changed significantly). Generally, VESPERV20 LST predictions are more accurate, for example in the Aral sea and Australia, indicating that V20 field set is informative and accurate. In contrast, there are some regions where the predictions get worse, for example at higher latitudes which is likely due to less accurate V20 field set over the area (i.e. lake location, lake mean depth values which influence lake freezing/thawing dates) and less valid Aqua-MODIS observations e.g. due to clouds. 7 points (two are overlaid in sub-Saharan Africa) where VESPERV20 LST prediction gets notably worse compared to VESPERV15 are highlighted with green circles and discussed in the text.



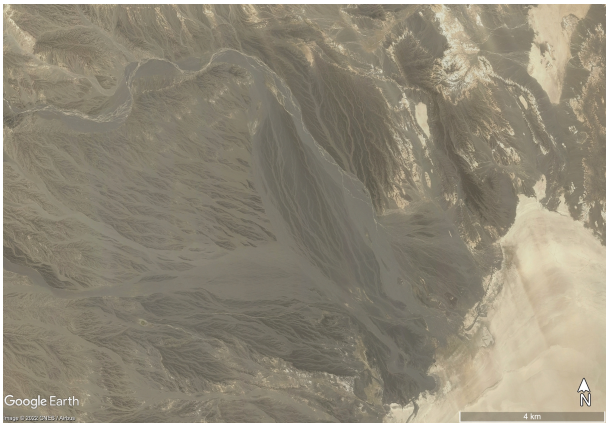
(a) Lake Natron, Tanzania



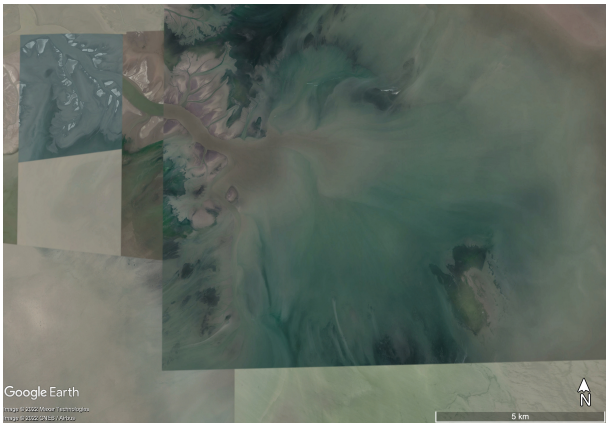
(b) Lake Blanche, Australia



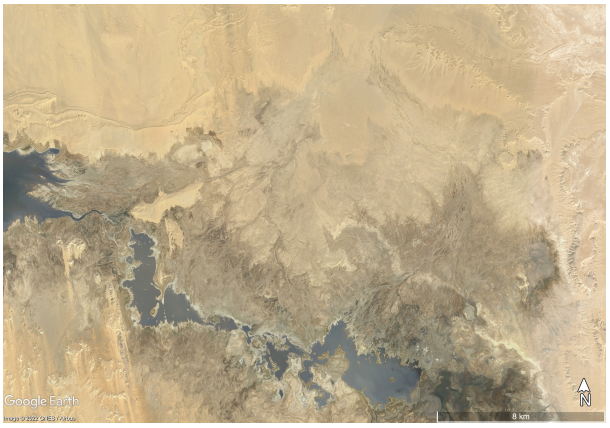
(c) Great Salt Lake Desert, Utah



(d) Farah Province, Afghanistan



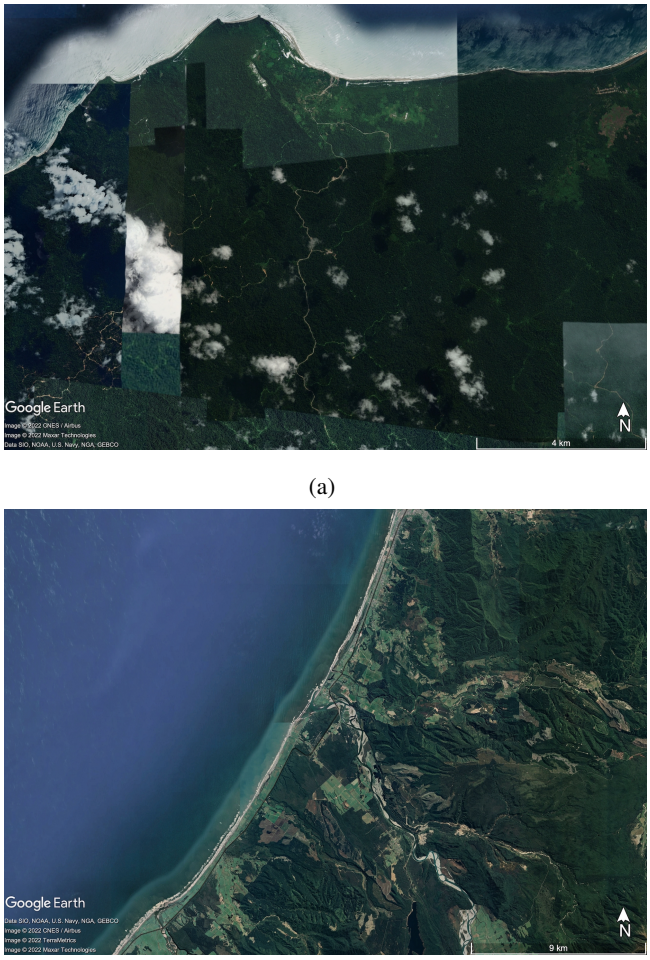
(e) Gujarat Province, India



(f) Toshka Lakes, Egypt

Figure 9. Satellite imagery of the problematic Lake Updates points highlighted in Fig. 8 where the V20 predictions are worse than the V15 predictions. Generally the updated V20 fields remove water, only considering permanent water. However these regions have highly time variable waters, which are better captured on average by the V15 fields. The images are centred on the grid box coordinates. Note that the lengthscales are different for some images.

comparing to VESPERV15 is linearly dependent to the forest reduction to bare ground in V20 field set comparing to V15 (i.e. strongest drop is when grid-cell fully covered with forest becomes fully covered with bare ground – high vegetation cover is reduced to zero). These erroneous grid-cells in V20 vegetation fields are likely to appear during the interpolation. The errors in these regions will in turn corrupt the LST predictions and mitigate the gain from a more accurate representation of the lake water.



(a)

(b)

Figure 10. Satellite imagery of grid boxes in *(a)* Siberut Island, Indonesia and *(b)* South Island, New Zealand. For both points it is expected according to the updated V20 fields that there is no vegetation, just bare ground. VESPER can identify these erroneous updated fields.

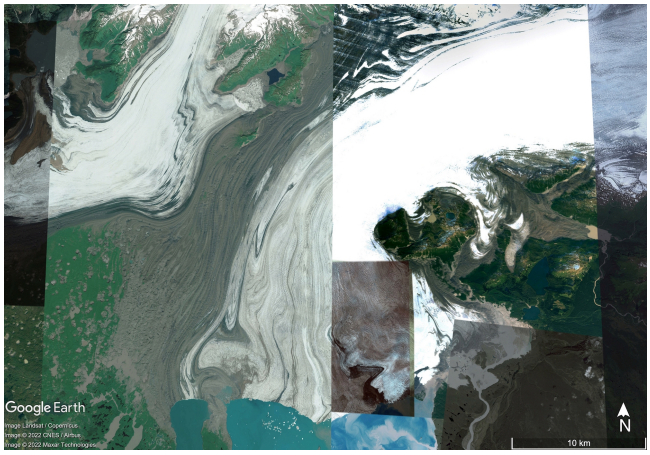
3.1.3 Category: Glacier Updates

Glacier Updates category in general shows improvement in LST predictability in VESPERV20 comparing to VESPERV15 (see Table 4) – prediction accuracy increases globally (over 1057 grid-cells) on average by 0.14K (training noise in VESPERV15

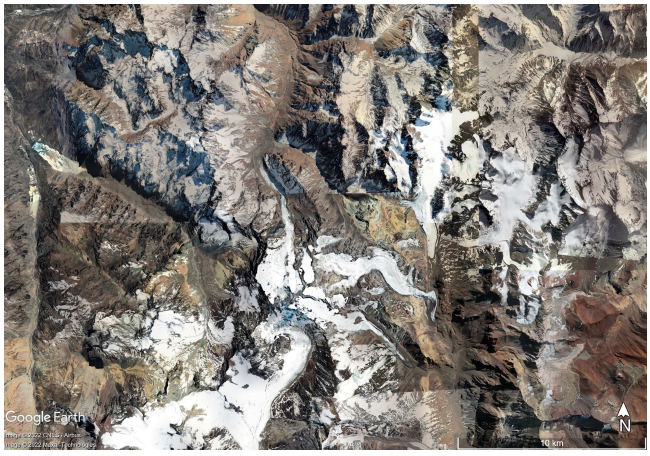
XX, and in VESPERV20 XX), most notably around the Himalayas, the land either side of the Davis strait, as well as British Columbia and the Alaskan Gulf. Analogous to the Lakes Updates category whilst the introduction of the V20 glacier cover generally improves LST predictions, there are some selected regions where the prediction gets worse. These are heavily concentrated in the southern hemisphere, in particular on the south-western edge of South America and the South Shetland Islands (which lie closer to Antarctica), and some parts of the Himalayas. This deterioration in performance in these areas is not due to erroneous update of V20 glacier cover, but related to the Aqua-MODIS data (i.e. sparse availability due to clouds, and less certain due to orography, see Figure 5). Consequently, VESPER finds it difficult to make accurate predictions in this region, and VESPERV20 settled into a local minimum which is worse than in VESPERV15 in these areas. If grid- cells with scarce amount of Aqua-MODIS observations (i.e. mean number of Aqua-MODIS observations per day over the year per ERA5 grid-cell is >50) are removed from the analysis then the worst performing grid-cells become excluded (yet a few areas where VESPERV20 underperforms VESPERV15 remain). 24 For example, there is a grid-cell in the Alaskan gulf on the Bering Glacier with δV20 = + 2.1 6K (training noise in VESPERV15 XX, and in VESPERV20 XX). This grid-cell has been updated in V20 field set comparing to V15 by strongly increasing glacier cover (from 0.68 to 0.92), decreasing low vegetation cover (from 0.10 to 0.01), modestly decreasing lake cover (from 0.07 to 0.01) and strongly increasing lake mean depth (from 2m to 27m). The Bering glacier is known to be a time varying (i.e. varies in size over the course of the season), and satellite imagery of the region (see Figure 10a) does give this impression; on longer timescales glacier exhibits a general reduction in its surface area (i.e. retreat of the terminus) over time, coupled with periodic surges in the glacier flow around every 20 years (Molnia and Post, 2010). It is therefore a complex region with even visible presence of low vegetation (see Figure 10a), so assuming it's constantly almost fully ice covered with no vegetation seems as an overcorrection. Another notable grid-cell is in the Chilean Andes, by the Juncal Glacier with δV20 = + 2.67K (training noise in VESPERV15 XX, and in VESPERV20 XX). Here V20 glacier cover was increased to 0.25 comparing to 0.00 in V15. According to the satellite imagery (see Figure 10b) the glacier only occupies a small fraction of the overall grid-cell, and the updated glacier cover may have been an over correction. Moreover, this is an complex orographic area with snowy mountain peaks at high altitude and deep valleys, therefore the temperature response due to the glacier feature could be atypical compared to e.g. the Alaskan Gulf or the Davis straight. For both of these points VESPER managed to identify inaccuracies in updated glacier cover, and once again proofed itself as a powerful tool for quality control of surface physiographic fields. From examples above it is evident that VESPER enables the user to quickly identify regions where the update to surface physiographic fields was beneficial (e.g. Aral Sea) and where it was not (e.g. Lake Natron, high vegetation updates). In turn, areas where LST predictions do not improve as expected can be inspected and erroneous or sub-optimal representations of the surface physiographic fields identified. This then provides key information on how and where to introduce additional corrections to better represent these more challenging or complex regions. Further some of these problematic areas are explored in more details and additional surface physiographic fields introduced with help of VESPER.

3.2 Evaluation of new lake fields: Monthly water & salt lakes

Particular regions where VESPER was struggling to make LST predictions – especially with the updated V20 field set which only include permanent water – were either areas with a large degree of temporal variability (e.g. lakes which flood and dry



(a)



(b)

Figure 11. Satellite imagery of *(a)* Bering Glacier, Alaskan Gulf, and *(b)* Juncal Glacier, Chile . For *(a)* it is expected that there is no vegetation and the grid box to be primarily ($\gtrsim 90\%$) dominated by ice. For *(b)* the updated V20 fields specify a $\gtrsim 25\%$ glacier fraction. Evidently, the V20 fields for these grid boxes are insufficiently accurate or informative, as identified by VESPER.

out periodically) or else areas with saline rather than freshwater lakes. Clearly if the size, shape and depth of a lake are changing over the course of the year, these are going to be hugely significant factors in modelling the lake temperature response. Similarly, saline lakes behave very differently to freshwater lakes since increased salt concentrations affect the density, specific heat capacity, thermal conductivity, and turbidity, as well as evaporation rates, ice formation and ultimately the surface temperature. These two properties of time variability and salinity are often related; it is common for saline lakes to flood and dry out over the course of the season, which naturally also affects the relative saline concentration of the lake itself. Currently, neither VESPERV15 or VESPERV20 have any information regarding the salinity of the lakes or their time variability. Indeed, FLake is specifically a fresh water lake model! This information can be introduced by including a global saline lake cover and

monthly varying lake cover as additional VESPER's input features, and use VESPER to rapidly assess accuracy of these new surface physiography fields and if their use in the model increase LST predictability. V20X field set is the same as V20 but with additional saline lake cover and difference between V20 static lake cover and monthly varying lake cover (12 maps in total); VESPERV20X is a neural network regression model trained on 2016 data with additional surface physiography fields, and used for LST predictions in 2019 (for comparison results with VESPERV15 see Table 4).

3.2.1 Category: Lake updates

For the Lake Updates category shows globally no significant difference in LST predictability if using V20X field set instead of V20 comparing to V15, for the Lake-Ground Updates category – LST predictability has slightly decreased (from V20=-1.12K to V20X=-1.09K, although the difference is rather small and within the training noise (see Table 4). Over problematic grid-cells discussed previously VESPERV20X LST predictability is significantly improved comparing to VESPERV20 (see Table 5).

Main surface physiographic fields (19 fields)	Pressure: surface pressure (sp, Pa), mean sea level pressure (msl, Pa), Wind: 10 metre U wind component (10u, m/s), 10 metre V wind component (10v, m/s), Temperature: 2 metre temperature (2t, K), 2 metre dewpoint temperature (2d, K), skin temperature (skt, K), ice temperature layer 1 (the sea-ice temperature in layer 0-7 cm; istl1, K), ice temperature layer 2 (the sea-ice temperature in layer 7-28 cm; istl2, K), Surface albedo: forecast albedo (fal, 0-1), Snow: snow depth (sd, m of water equivalent)
Scenarie:	At oprette en server med bestemte regler som tillader folk at spille sammen. More Text more text More Text

Table 4. *TK: This table is a placeholder for Table 5 in the word doc*

For Lake Blanche, use of V20X field set reduces the LST prediction error by 2.43K compared to use of V20. This is inspite of the fact that current saline lake cover do not identify Lake Blanche as a salt lake, and so all the improvement in the prediction is from the additional information from the monthly lake maps. The salt lake maps are similarly inaccurate for the grid point in Northern India, failing to recognise the underlying salt marsh. However, again the information contained in the monthly lake maps allows the reduction in the error by 2.19K. There are also regions where the saline maps are correct to not specify any salinity, such as in Afghanistan and the Toshka lakes; again the monthly maps provided sufficient information to allow for a marked improvement by 2.04 K and 2.15 K respectively. This is particularly notable since the size of the monthly lake corrections is small for these points: the mean monthly correction for Afghanistan is 0.046 and for the Toshka Lakes is 0.001. However, under the updated V20 both of these areas have had all water removed and so adding in just a small amount of time variable water allows for much more accurate predictions. This example illustrates how VESPER both can identify inaccurate fields and quantify the value of updated fields, as well as emphasizing the importance of time variable lake fields more generally. For points where the saline maps do specify that the underlying lakes are salt lakes - Lake Natron and Salt Lake Desert - it is not possible to disentangle whether the gain is due to the saline maps or the monthly maps. The centre of Lake Natron

exhibits a particularly notable improvement by 2.6K, whilst for the grid point on the northern edge the gain is more modest at 0.12K. This is likely due to the fact that the updated monthly maps provided much stronger corrections at the centre of the lake (mean correction 0.13) than at the northern edge (mean correction 0.02). For the grid point at the Great Salt Lake Desert, the improvement is 2.06 K again with a strong correction from the monthly lake maps (mean value 0.16) and the salt maps (mean value 0.56). This improvement is to be expected given the known strong salinity and time variability in the region, and so it is a nice confirmation to have these updated fields verified by VESPER. It is also notable that the variation in the monthly lake maps at this point is very large, with a standard deviation in the lake fraction over 12 months of 0.18. At the start of the year the corrections from the monthly maps are very large, then as the year progresses the magnitude of the corrections generally decreases as the lake dries out. Such a large variation is again difficult to ever capture with a static field. Other regions of note that we have mentioned previously are the Northwest Territories and the Nunavut province in Northern Canada where the V20 model underperformed relative to V15, with δV20 = +0.02 K. The introduction of the monthly lake maps modestly improves the predictions in this area, with δV20X = -0.03K. In these high latitude regions one might expect some time variability due to freezing and thawing of the lake surfaces, and the addition of the monthly lake maps to the model then provides some of this time variable information, allowing for improved predictions. Whilst this is an improvement, the effect is modest; it is generally difficult to get quality observations at high latitudes, especially during the cold season, due to increased cloud cover. Therefore whilst VESPER can say that the addition of the monthly lake maps does improves the predictions in these regions, for improved performance cloud independent data should be used. Additionally, the corrections from the monthly lake maps are small in these regions, with a mean correction of 0.02 and a generally small variance; in actuality time variable fields with greater variance over the year may be more accurate. Due to the freezing and thawing, improving ice on/off date prediction by the lake parametrisation should help better describe the seasonality and variance. It is worth emphasising that whilst the V20 and V20X models are improvements over V15 globally, and V20X is generally an improvement over V20 for these problematic points, there are regions where neither V20 or V20X outperform V15 (δM is always positive), such as Lake Natron and Northern India. Given all the extra information provided to the more advanced models this is unusual, unless the additional information is erroneous in these regions or else the temperature response is completely atypical to the rest of the globe and the additional information is not predictive in these regions. To explore this hypotheses we train one further model, V15X. This is analogous to V20X, being simply the V15 model with the additional monthly maps and salt lake fields included. Importantly it does not have the updated correction fields from V20. Globally, this model performs worse that the V20 models, as we might expect - for example in the Lake Updates category δV15X = -0.25K compared to δ20X = -0.45 K. However, V15X does perform well at these problematic points (see Table 3). For both the Lake Natron grid points V15 outperforms V20X, suggesting that at this location the V20 fields are generally less accurate than the V15 fields. V15X however underperforms relative to V15 which also indicates that the monthly maps and the salt lakes are either inaccurate at this location, or that the temperature response of Lake Natron is highly atypical. For Northern India, the performance of the V15 model is particularly striking; whilst the V20 and V20X models struggled to make more accurate predictions than V15, V15X decreases the average prediction error by nearly 6K. This again indicates that for this point the V20 fields are less accurate than V15. Similarly for the Great Salt Lake Desert, δV20 = 1.78 K, δV20X = -0.28 K but δV15X = -0.86 K, which suggests that whilst the monthly lake maps and

the salt lake fractions are accurate and informative in this area, the static V20 fields are not. These examples illustrates again how VESPER can identify particular regions where the fields are inaccurate, as well as emphasising the need more generally for accurate descriptions of seasonally varying inland water and saline lake maps in Earth system modelling

3.2.2 Category: Vegetation Updates

Whilst the Vegetation Updates category explicitly deals with areas where the lake fraction does not change when going from V15 to V20, many of the grid points in this category do contain some kind of waterbody, often lying close to the coast or else containing lakes or large rivers. Information on the salinity and temporal variability of these water bodies can then influence the prediction accuracy. By providing the additional information in V20X, the mean δM is reduced from δV20 = +0.049K to δV20X = 0.005K. This performance is gained despite the known errors for some of the grid boxes in the cvh vegetation updates (e.g. Figure 7), again demonstrating the importance of salinity and seasonally varying water. The V15X model is less performant than V20X, with δV15X = 0.11K since there are some grid boxes in this category where the updated V20 fields are accurate and valuable if augmented by monthly variability. However if we consider just the worst performing grid points where δV20 > 1 K then the mean values are δV20 = 2.0 K, δV20X = 0.49 K and δV15X = 0.21K. This again demonstrates how the cvh fields have been erroneously updated for a small selection of grid points in V20.

3.2.3 Category: Glacier Updates

We would expect the additional information provided by V20X to be particularly effective for glacial grid points. Glacier ice is naturally found next to waterbodies which freeze and thaw over the year, and the salinity of water will also influence this freezing. Therefore accurate additional information from the monthly lake maps and the saline maps should prove useful in these more time variable regions. This is indeed what we observe with the mean delta going from δV20 = -0.13K to δV20X = -0.24K. Considering the two problematic points that we discussed previously, in the Alaskan Gulf the prediction accuracy relative to V15 has improved from δV20 = +2.16 K to δV20X = +1.00 K, whilst for the Juncal Glacier the prediction accuracy has also improved, with δV20X decreasing to +1.88 K. Despite this improvement, again for both of these points the prediction accuracy still lags behind V15. This is on account of the V20 fields being insufficiently accurate in these areas, as has been discussed. Neither of these grid points correspond to saline lakes or have a significant time variability in the monthly lake fractions and so are also not improved by a V15X model.

3.2.4 Timeseries

Thus far we have been focusing mainly on the δM metric averaged over the entire year of the test set. It is also of interest to explore how the prediction error for each of the 3 models varies with time. This is demonstrated in Fig 9 for each of the 4 updated categories that we have discussed. For the Lake Updates and Lake-Ground Updates categories we can see that all the model predictions track the same general profile, with the error peaking in the northern hemisphere summer months. This is a result of FLake modelling being least accurate during the summer as the lake is not fully mixed and so the mixed layer

depth for lakes is too shallow, resulting in skin temperatures with larger errors. Conversely, in the autumn and spring the lake is fully mixed and predictions have the smallest errors compared with observations. A clear hierarchy of models is evident; the V20/V20X models consistently outperform the V15 model across the year. This again is solid evidence, highlighted by VESPER, of the value of the updated fields both static and seasonally varying. We mentioned previously how the annually and globally averaged δM values for the Lake Updates category were highly comparable for V20 and V20X, despite V20X significantly improving the worst behaving points. We can see from the top panel in Figure 9 that the V20X model is a systematic improvement on V20 from around April onwards, but at earlier times in the year V20 outperforms V20X. This is likely for two reasons. Firstly, the monthly lake maps are in fact a climatology and therefore insufficiently precise to detect the exact ice on/off dates during the winter months, where we have a large number of grid points at high latitudes which will be subject to freezing, nullifying any time variability. Secondly, due to the accuracy of the lake mean depth which strongly drives the ice on-dates due to its influence on the heat capacity of the lake. During the warmer months lakes thaw, the monthly maps are more accurate, as the thawing of lake ice is mainly connected to the atmospheric conditions, not the lake depth, and so the information contained in them can be used to make more accurate predictions. The Lake-Ground Updates timeseries broadly follows the same general profile as Lake Updates, but the errors are larger - those grid points where the lakes have been replaced with bare ground were particularly poorly described in V15. Additionally, for Lake Updates we see two sharp decreases in the prediction error during $\sim$ April and September, which are not as strongly reflected in Lake-Ground. This is due to the geographic location of the grid points in each of the two categories; for the Lake Updates category the grid points are located primarily in the boreal zones and so are subject to freezing and thawing over the course of the year leading to a strong seasonality due to the lake mixing that we have discussed. The sharp drop in April corresponds to a time where the lakes are unfrozen and fully mixed. However the lakes in the Lake-Ground sub-category are less concentrated and much more evenly distributed over the globe and so do not exhibit such a strong seasonality.

Examining the timeseries for Vegetation Updates, whilst there is a large degree of variability, we again see the trend previously discussed whereby the V20 fields make the predictions systematically worse across the entirety of the test year. The introduction of the monthly lake maps in V20X compensates for the erroneous V20 vegetation fields - the V20X predictions are generally worse than V15 at the start and end of the year, but better in the middle. This is due to the fact that the majority of the points in the Vegetation Updates category are in climate zones which have pronounced rainy and dry seasons e.g. Indonesia, the Amazon. At the start of the year during the wet season there is lots of precipitation and the static V15 fields are generally more accurate. As the rains abate and the dry season starts the V15 lake fields are underestimates which are then improved through the introduction of additional water via the monthly lake maps. For Glacier updates we can separate grid points into the northern and southern hemispheres. We again consider just those points where the number of MODIS observations at a given instant in time, per ERA data point is greater than 50. For the northern hemisphere the familiar hierarchy of models is recovered, with the V20X model generally outperforming V20, which in turn generally outperforms V15. The errors peak for all models in the summer, again due to the lakes not being fully mixed. There is also an uptick in the prediction error for all models during the winter when the freezing is greatest - this indicates how ice cover can strongly influence the land surface temperature response. For the southern hemisphere the story is different. The errors are smallest during the middle of the year

when we expect the freezing to be greatest. During the spring and autumn the errors are largest - this is correlated with a decrease in the number of observations suggesting that this is due to poorer data quality due to cloud cover. In the summer when the weather is clearer the errors start to decrease again. Given this variation in the data quality due to cloud cover it is difficult to draw any strong conclusions, and again for stronger performance cloud independent data should be used. What is obvious for the southern hemisphere glacier grid points is that the V20 and V20X models struggle to improve on V15, unlike in the northern hemisphere. This suggests that the updated V20 fields are still insufficiently accurate for southern latitudes. We have also discussed previously particular grid points that will likely show a large degree of temporal variability or the lakes are saline and as a consequence the static V15/V20 fields struggle to make accurate predictions (e.g. Table 3). In Figure 10 we present timeseries for two of these points: Lake Natron in Tanzania and Gujarat Province, India. Both these points were discussed in Sections 3.1.1 and 3.2.1. We can see that for these two selected points the hierarchy of models no longer holds. Whilst there is a large degree of variability, and there is no clear separation between models that we get when averaging over all grid points as in Fig 9, generally it can be seen that the V20 model performs the worst, indicating that the updated fields are not accurate in these regions. For Lake Natron the V20/V20X models are significantly worse throughout almost the entire year. The updated models - which specify a much larger lake fraction than in V15 - perform well at the beginning and start of the year which tend to be the wettest months at Lake Natron. However, during the summer as the lake dries out the errors grow significantly. This indicates again that the updated V20 fields are in fact over-corrections for this area. Similarly, whilst the V15X model is a significant improvement over V20/V20X - since it does not have these inaccurate fields - it is still less performant than the basic V15 again due to the additional water that V15X specifies. Together this strongly indicates that there is little surface water at Lake Natron during 2019.

For Gujarat Province in Northern India the story is different. Now the V20 model is systematically worse than V15 over the entire year, indicating that the static V20 fields are less accurate than the V15 fields. The V20X model shows a strong time variability, with the errors being smallest in the summer which is the wet season in Gujarat and largest in the winter which is the dry season. This suggests that the monthly maps are most accurate during the summer, providing extra information which is missing from V15/V20, but may be overestimates during the winter. The V15X model has a notably strong performance, outperforming the other models almost every month. This again is further evidence of the inaccuracy of the V20 fields and the value of the time-variable monthly water information.

4 Discussion

We have seen how VESPER can quantitatively evaluate the value of updates to the lake surface parametrisation as well as identifying areas where the updates are inaccurate. For the former VESPER was able to show that the major regions where the lake surface parametrisation fields were updated - such as the Aral sea - enjoyed more accurate predictions, which verifies both the accuracy of the fields and their information content with respect to predicting skin temperatures. For the latter VESPER was able to identify grid points where the predictions became worse with the updated fields, indicating that the updated fields were in fact less accurate. More generally we have also seen how detailed knowledge of surface water fields (e.g. up to date

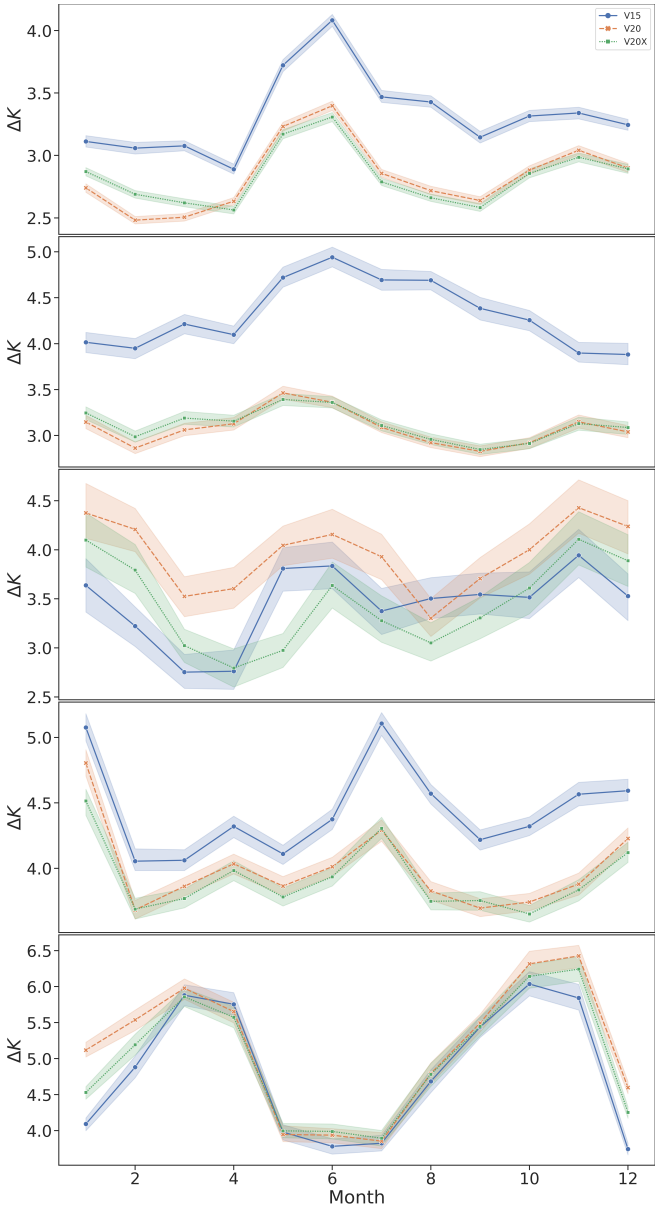


Figure 12. Mean prediction error in the surface temperature $\bar{\Delta}K$, averaged over all grid points, for each of the 3 models over the course of the test year for (*top panel*) Lake Updates, (*second panel*) Lake-Ground Updates, (*third panel*) Vegetation Updates, (*fourth panel*) Glacier Updates, northern hemisphere and (*bottom panel*) Glacier Updates, southern hemisphere. For the Glacier Updates category we again exclude grid points where the number of MODIS observations per ERA data point is less than 50. For the Lake categories, all models follow the same general profile, with the V20X model generally outperforming the V20 model over the year, which in turn outperforms the V15 model. The value of the additional V20 correction fields and the V20X monthly lake maps and salt lake maps, is evident.

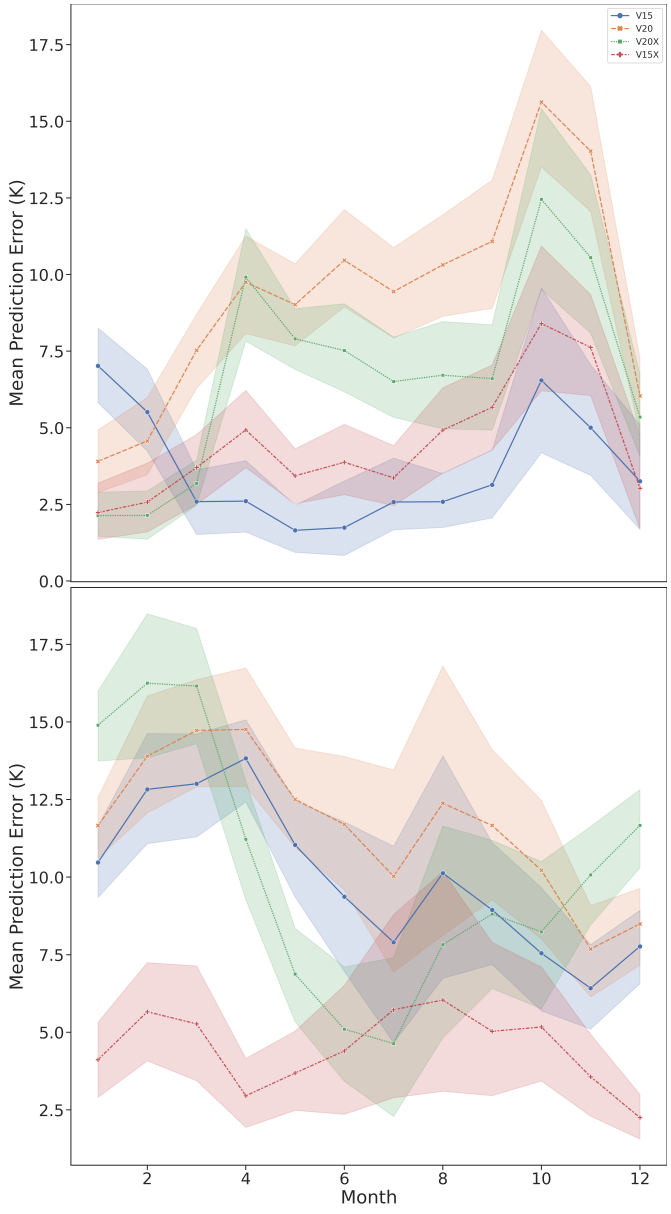


Figure 13. Variation in the prediction error for the grid points at Lake Natron, Tanzania (top panel) and Gujarat Province, India (bottom panel). There is a large degree of variability, but for Lake Natron the V20 and V20X models are generally less performant than V15 and V15X, indicating that the updated V20 fields are less accurate here. The augmented models with saline and monthly lake maps outperform those without, indicating the value of these fields in these regions.

permanent water distribution, seasonal water distribution, salt lake distribution, etc.) can notably improve the accuracy with which the skin temperature can be modelled, e.g. grid points with significant updates (i.e. where the field has changed by GRT EQ 10 %) to the lake fields show a mean absolute error reduction of skin temperature globally of 0.45K (Table 2). There are multiple possible further extensions of this work. We have not currently included the errors on the MODIS observations into the VESPER model. During the “matching-in-space" step relating the ERA and MODIS data (Section 2.2), it could be a worthwhile extension to weight the averaged MODIS points by their corresponding errors (e.g. Fig. 2) when deriving a single MODIS observation for a given ERA grid point. This would then provide a more accurate and confident representation of the true surface temperature at a particular space-time point. Due to the inherent stochasticity of training a model it is also possible for different models to settle in different local minimas i.e. the network variance. It would also be desirable to train an ensemble of models (“ensemble learning") and combine the predictions from multiple models to reduce this variance. We have focused here primarily on hydrological applications, our primary concern being the ability to evaluate the parametrised water body representation, however the method would work generally for any updated fields that we want to assess. Extension to non-lake hydrological fields like wetland extent or river bathymetry model parameters, or even non hydrological fields such as orography would be an interesting further development. The development of a more mature, integrated pipeline for automatically evaluating updated parametrisations could also be a worthwhile pursuit. Another natural extension of this work which may prove fruitful in the enterprise for improved parametrised representation of water bodies is to invert the problem and treat VESPER as a function to optimise. That is to say, VESPER can be thought of as a function which takes some inputs - in this case a lake parametrisation - and returns a loss metric i.e. how accurate the predictions are compared to the test set. Given this loss metric it may then be possible to vary the inputs and use standard optimisation techniques to learn the optimal parametrisation. Whilst this may be an expensive technique as there are effectively two nested models over which to optimise (for every optimisation step in the higher model, one must train the VESPER network from scratch) it could be possible given appropriate hardware or with reduced data focusing just on targeted locations (e.g. “What is the best way to represent the lakes in this area?"). The loss gradient information can also be used to tune individual features, informing whether an input variable should be larger or smaller.

5 Conclusion

Weather and climate modelling rely on accurate, up-to-date descriptions of surface fields, such as inland water, so as to provide appropriate boundary conditions for the numerical evolution. Lakes can significantly influence both weather and climate, but sufficiently accurate representation of lakes is challenging and the natural changes in water bodies mean that these representations need to be frequently updated. A new method based on a neural network regressor for automatically and quickly verifying the updated lake fields - VESPER - has been presented in this work. This tool has been deployed to verify the recent updates to the FLake parametrisation, which include additional datasets such as the GSWE and updated methods for determining the lake depth from GLDBv3. The updated parametrisation fields were shown globally to be an improvement over the original fields; for a subset of grid points which have had significant updates to the lake fields, the prediction error in the skin temperature de-

creased by 0.45K. Conversely, VESPER also identified individual grid points where the updated lake fields were less accurate, enabling these points to subsequently be corrected, such as losing forests to bare ground leading to errors of 1.1K. Multiple further extensions of this work, including extension to non lake fields and the development of a more mature integrated pipeline have been discussed.

Code availability. The code used in constructing VESPER, including the methods for joining the ERA and MODIS datasets and the construction of the neural network regression model is open-sourced at https://github.com/tomkimpson/ML4L

Author contributions. All the authors contributed equally to the work.Tom Kimpson wrote the manuscript with contributions from all other authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant No 741112).

References

Two Decades of Change at Toshka Lakes, https://earthobservatory.nasa.gov/images/149334/two-decades-of-change-at-toshka-lakes, accessed: 2022-09-30.

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, https://www.tensorflow.org/, software available from tensorflow.org, 2015.

Arino, O., Ramos Perez, J. J., Kalogirou, V., Bontemps, S., Defourny, P., and Van Bogaert, E.: Global Land Cover Map for 2009 (GlobCover 2009), https://doi.org/10.1594/PANGAEA.787668, 2012.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., and Lindauer, M.: Hyperparameter Optimization: Foundations, Algorithms, Best Practices and Open Challenges, arXiv e-prints, arXiv:2107.05847, 2021.

Bontemps, S., Defourny, P., Van Bogaert, E., Arino, O., Kalogirou, V., and Ramos Perez, J.: GLOBCOVER 2009 Product description and validation report, http://due.esrin.esa.int/page_globcover.php, accessed: 2022-06-01, 2011.

Boussetta, S., Balsamo, G., Arduini, G., Dutra, E., McNorton, J., Choulga, M., Agustí-Panareda, A., Beljaars, A., Wedi, N., Munõz-Sabater, J., de Rosnay, P., Sandu, I., Hadade, I., Carver, G., Mazzetti, C., Prudhomme, C., Yamazaki, D., and Zsoter, E.: ECLand: The ECMWF Land Surface Modelling System, Atmosphere, 12, https://doi.org/10.3390/atmos12060723, 2021.

Center, N. G. S. F.: MODIS, https://modis.gsfc.nasa.gov/, accessed: 2022-06-01.

Chantry, M., Hatfield, S., Duben, P., Polichtchouk, I., and Palmer, T.: Machine learning emulation of gravity wave drag in numerical weather forecasting, in: EGU General Assembly Conference Abstracts, EGU General Assembly Conference Abstracts, pp. EGU21–7678, https://doi.org/10.5194/egusphere-egu21-7678, 2021.

Choulga, M., Kourzeneva, E., Zakharova, E., and Doganovsky, A.: Estimation of the mean depth of boreal lakes for use in numerical weather prediction and climate modelling, Tellus A: Dynamic Meteorology and Oceanography, 66, 21 295, https://doi.org/10.3402/tellusa.v66.21295, 2014.

Choulga, M., Kourzeneva, E., Balsamo, G., Boussetta, S., and Wedi, N.: Upgraded global mapping information for earth system modelling: an application to surface water depth at the ECMWF, Hydrology and Earth System Sciences, 23, 4051–4076, https://doi.org/10.5194/hess-23-4051-2019, 2019.

DelSontro, T., Beaulieu, J. J., and Downing, J. A.: Greenhouse gas emissions from lakes and impoundments: Upscaling in the face of global change, Limnology and Oceanography Letters, 3, 64–75, https://doi.org/https://doi.org/10.1002/lol2.10073, 2018.

Düben, P., Modigliani, U., Geer, A., Siemen, S., Pappenberger, F., Bauer, P., Brown, A., Palkovic, M., Raoult, B., Wedi, N., and Baousis, V.: Machine learning at ECMWF: A roadmap for the next 10 years, https://doi.org/10.21957/ge7ckgm, 2021.

Eerola, K., Rontu, L., Kourzeneva, E., Pour, H. K., and Duguay, C.: Impact of partly ice-free Lake Ladoga on temperature and cloudiness in an anticyclonic winter situation – a case study using a limited area model, Tellus A: Dynamic Meteorology and Oceanography, 66, 23 929, https://doi.org/10.3402/tellusa.v66.23929, 2014.

Franz, D., Mammarella, I., Boike, J., Kirillin, G., Vesala, T., Bornemann, N., Larmanou, E., Langer, M., and Sachs, T.: Lake-Atmosphere Heat Flux Dynamics of a Thermokarst Lake in Arctic Siberia, Journal of Geophysical Research: Atmospheres, 123, 5222–5239, https://doi.org/https://doi.org/10.1029/2017JD027751, 2018.

Freire, S., MacManus, K., Pesaresi, M., Doxsey-Whitfield, E., and Mills, J.: Development of new open and free multi-temporal global population grids at 250 m resolution, AGILE, https://agile-online.org/Conference_Paper/cds/agile_2016/shortpapers/152_Paper_in_PDF.pdf, 2016.

Hatfield, S., Chantry, M., Dueben, P., Lopez, P., Geer, A., and Palmer, T.: Building Tangent-Linear and Adjoint Models for Data Assimilation With Neural Networks, Journal of Advances in Modeling Earth Systems, 13, e02521, https://doi.org/10.1029/2021MS002521, 2021.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/https://doi.org/10.1002/qj.3803, 2020.

Hewson, T. D. and Pillosu, F. M.: A low-cost post-processing technique improves weather forecasts around the world, Communications Earth and Environment, 2, 132, https://doi.org/10.1038/s43247-021-00185-9, 2021.

Huang, W., Cheng, B., Zhang, J., Zhang, Z., Vihma, T., Li, Z., and Niu, F.: Modeling experiments on seasonal lake ice mass and energy balance in the Qinghai–Tibet Plateau: a case study, Hydrology and Earth System Sciences, 23, 2173–2186, https://doi.org/10.5194/hess-23-2173-2019, 2019.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, arXiv e-prints, arXiv:1412.6980, 2014.

Kourzeneva, E., Asensio, H., Martin, E., and Faroux, S.: Global gridded dataset of lake coverage and lake depth for use in numerical weather prediction and climate modelling, Tellus A: Dynamic Meteorology and Oceanography, 64, 15 640, https://doi.org/10.3402/tellusa.v64i0.15640, 2012.

Liu, H., Jezek, K., Li, B., and Zhao, Z.: Radarsat Antarctic Mapping Project Digital Elevation Model, Version 2, https://doi.org/10.5067/8JKNEW6BFRVD, 2015.

lu, P., Cao, X., Li, G., Huang, W., Leppäranta, M., Arvola, L., Huotari, J., and Li, Z.: Mass and Heat Balance of a Lake Ice Cover in the Central Asian Arid Climate Zone, Water, 12, 2888, https://doi.org/10.3390/w12102888, 2020.

Mironov, D. V.: Parameterization of lakes in numerical weather prediction: Description of a lake model, DWD Offenbach, Germany, 2008.

Molnia, B. F. and Post, A.: Surges of the Bering Glacier, in: Bering Glacier: Interdisciplinary Studies of Earth's Largest Temperate Surging Glacier, Geological Society of America, https://doi.org/10.1130/2010.2462(15), 2010.

Muñoz Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth System Science Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

Munoz Sabater, J.: ERA5-Land hourly data from 1981 to present., https://doi.org/10.24381/cds.e2161bac, 2019.

Notaro, M., Zarrin, A., Vavrus, S., and Bennington, V.: Simulation of Heavy Lake-Effect Snowstorms across the Great Lakes Basin by RegCM4: Synoptic Climatology and Variability, Monthly Weather Review, 141, 1990 – 2014, https://doi.org/10.1175/MWR-D-11-00369.1, 2013.

Pace, M. and Prairie, Y.: Respiration in lakes, Respiration in Aquatic Ecosystems, https://doi.org/10.1093/acprof:oso/9780198527084.003.0007, 2005.

Parkinson, C.: Aqua: An Earth-Observing Satellite Mission to Examine Water and Other Climate Variables, Geoscience and Remote Sensing, IEEE Transactions on, 41, 173 – 183, https://doi.org/10.1109/TGRS.2002.808319, 2003.

Pekel, J.-F., Cottam, A., Gorelick, N., and Belward, A.: High-resolution mapping of global surface water and its long-term changes, Nature, 540, https://doi.org/10.1038/nature20584, 2016.

RAPIDS: cuML - GPU Machine Learning Algorithms, https://github.com/rapidsai/cuml, v22.04.00.

Saunois, M., Stavert, A. R., Poulter, B., Bousquet, P., Canadell, J. G., Jackson, R. B., Raymond, P. A., Dlugokencky, E. J., Houweling, S., Patra, P. K., Ciais, P., Arora, V. K., Bastviken, D., Bergamaschi, P., Blake, D. R., Brailsford, G., Bruhwiler, L., Carlson, K. M., Carrol, M., Castaldi, S., Chandra, N., Crevoisier, C., Crill, P. M., Covey, K., Curry, C. L., Etiope, G., Frankenberg, C., Gedney, N., Hegglin, M. I., Höglund-Isaksson, L., Hugelius, G., Ishizawa, M., Ito, A., Janssens-Maenhout, G., Jensen, K. M., Joos, F., Kleinen, T., Krummel, P. B., Langenfelds, R. L., Laruelle, G. G., Liu, L., Machida, T., Maksyutov, S., McDonald, K. C., McNorton, J., Miller, P. A., Melton, J. R., Morino, I., Müller, J., Murguia-Flores, F., Naik, V., Niwa, Y., Noce, S., O'Doherty, S., Parker, R. J., Peng, C., Peng, S., Peters, G. P., Prigent, C., Prinn, R., Ramonet, M., Regnier, P., Riley, W. J., Rosentreter, J. A., Segers, A., Simpson, I. J., Shi, H., Smith, S. J., Steele, L. P., Thornton, B. F., Tian, H., Tohjima, Y., Tubiello, F. N., Tsuruta, A., Viovy, N., Voulgarakis, A., Weber, T. S., van Weele, M., van der Werf, G. R., Weiss, R. F., Worthy, D., Wunch, D., Yin, Y., Yoshida, Y., Zhang, W., Zhang, Z., Zhao, Y., Zheng, B., Zhu, Q., Zhu, Q., and Zhuang, Q.: The Global Methane Budget 2000–2017, Earth System Science Data, 12, 1561–1623, https://doi.org/10.5194/essd-12-1561-2020, 2020.

Schiavina, M., Freire, S., and MacManus, K.: GHS-POP R2022A - GHS population grid multitemporal (1975-2030), https://ghsl.jrc.ec.europa.eu/ghs_pop2022.php, 2022.

Thiery, W., Davin, E. L., Panitz, H.-J., Demuzere, M., Lhermitte, S., and van Lipzig", N.: The Impact of the African Great Lakes on the Regional Climate, Journal of Climate, 28, 4061 – 4085, https://doi.org/10.1175/JCLI-D-14-00565.1, 2015.

Thiery, W., Gudmundsson, L., Bedka, K., Semazzi, F. H. M., Lhermitte, S., Willems, P., van Lipzig, N. P. M., and Seneviratne, S. I.: Early warnings of hazardous thunderstorms over Lake Victoria, Environmental Research Letters, 12, 074 012, https://doi.org/10.1088/1748-9326/aa7521, 2017.

Tranvik, L. J., Downing, J. A., Cotner, J. B., Loiselle, S. A., Striegl, R. G., Ballatore, T. J., Dillon, P., Finlay, K., Fortino, K., Knoll, L. B., Kortelainen, P. L., Kutser, T., Larsen, S., Laurion, I., Leech, D. M., McCallister, S. L., McKnight, D. M., Melack, J. M., Overholt, E., Porter, J. A., Prairie, Y., Renwick, W. H., Roland, F., Sherman, B. S., Schindler, D. W., Sobek, S., Tremblay, A., Vanni, M. J., Verschoor, A. M., von Wachenfeldt, E., and Weyhenmeyer, G. A.: Lakes and reservoirs as regulators of carbon cycling and climate, Limnology and Oceanography, 54, 2298–2314, https://doi.org/https://doi.org/10.4319/lo.2009.54.6_part_2.2298, 2009.

Vavrus, S., Notaro, M., and Zarrin, A.: The Role of Ice Cover in Heavy Lake-Effect Snowstorms over the Great Lakes Basin as Simulated by RegCM4, Monthly Weather Review, 141, 148 – 165, https://doi.org/10.1175/MWR-D-12-00107.1, 2013.

Verpoorter, C., Kutser, T., Seekell, D. A., and Tranvik, L. J.: A global inventory of lakes based on high-resolution satellite imagery, Geophysical Research Letters, 41, 6396–6402, https://doi.org/https://doi.org/10.1002/2014GL060641, 2014.

Viterbo, P.: A review of parametrization schemes for land surface processes, https://www.ecmwf.int/node/16960, 2002.

Wan, Z.: MODIS Collection 6.1 (C61) Product User Guide, https://lpdaac.usgs.gov/documents/715/MOD11_User_Guide_V61.pdf, accessed: 2022-06-01, 2019.

Wan, Z., Hook, S., and Hulley, G.: MODIS/Aqua Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061 [Data set], https://doi.org/10.5067/MODIS/MYD11A1.06, accessed: 2022-06-01.

Yu, T. and Zhu, H.: Hyper-Parameter Optimization: A Review of Algorithms and Applications, arXiv e-prints, arXiv:2003.05689, 2020.