Talajtani projekt az ELKH Cloud infrastruktúrán – tapasztalatok megosztása

Tóth (Szabó) Brigitta

Agrártudományi Kutatóközpont, Talajtani Intézet

ELKH
Eötvös Loránd
Kutatási Hálózat

ATK Agrártudományi Kutatóközpont

euptf_v2 projekt

- Projekt igénylése: 2018. augusztus

Bolyai ösztöndíjas kutatás

Az európai talajokra kidolgozott talajhidrológiai becslő eljárások tovább pontosíthatók a legújabb adatbányászati módszerek alkalmazásával. A módszerek pontossága mellett fontos szempont azok nagy adatokon történő alkalmazhatósága és a gyakorlat számára is fontos információk kinyerésének lehetősége.

1. Adatbányászati módszerek áttekintése hangsúlyt fektetve az optimalizálás, párhuzamos számítás és interpretáció lehetőségeire.

2. A kiválasztott adatbányászati eljárások hatékonyságának és értelmezhetőségének vizsgálata az európai talajhidrológiai adatbázison (EU-HYDI).

3. Az európai talajhidrológiai becslő módszerek (Tóth et al., 2015: https://onlinelibrary.wiley.com/doi/abs/10.1111/ejss.12192 , https://esdac.jrc.ec.europa.eu/taxonomy/term/10) becslési hatékonyságának növelése adatbányászati eljárás alkalmazásával.

Eredmények publikálása nemzetközi impakt faktoros folyóiratban.

- eduID
- Belépéshez szükséges adatok e-mailben és sms-ben
- Infrastruktúra definiálása (virtuális gép)
- Hosszabbítás

Ha a projekt véget ért

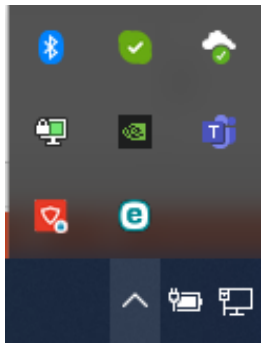
Ha sikeresen csatlakozott és használta az ELKH Cloud erőforrásait és ennek alapján cikket ír, kérjük, hogy cikkében a következő magyar, vagy angol nyelvű köszönetnyilvánítást helyezze el:

> Köszönetnyilvánítás
> A ..... projekt nevében köszönetet mondunk az ELKH Cloud (https://science-cloud.hu/) használatáért, ami nagyban hozzájárult a publikált eredmények eléréséhez.
>
> Acknowledgement
> On behalf of Project ..... we thank for the usage of ELKH Cloud (https://science-cloud.hu/) that significantly helped us achieving the results published in this paper.

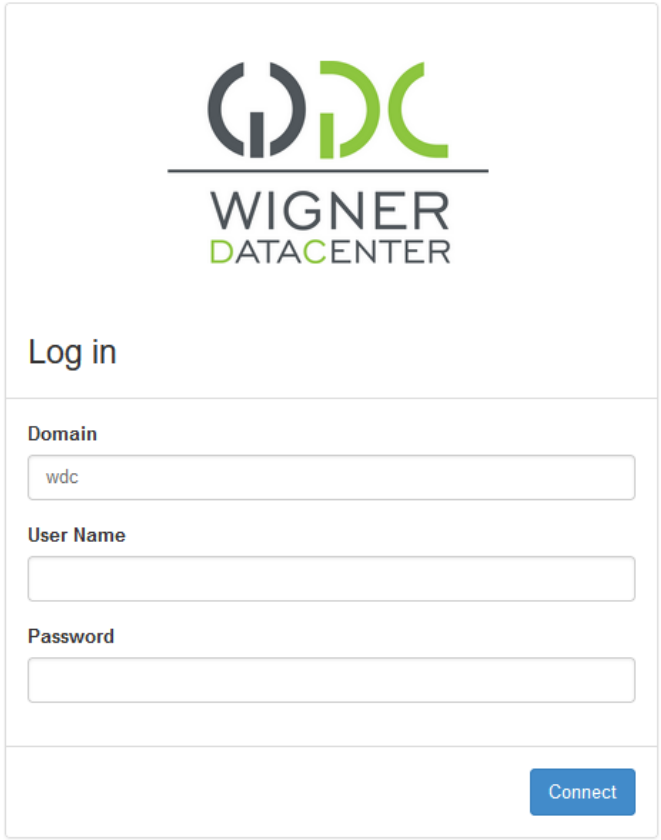
Megjelent cikk és hivatkozási paraméterek küldése ELKH Cloud Csapatának.

Belépés a WDC-be

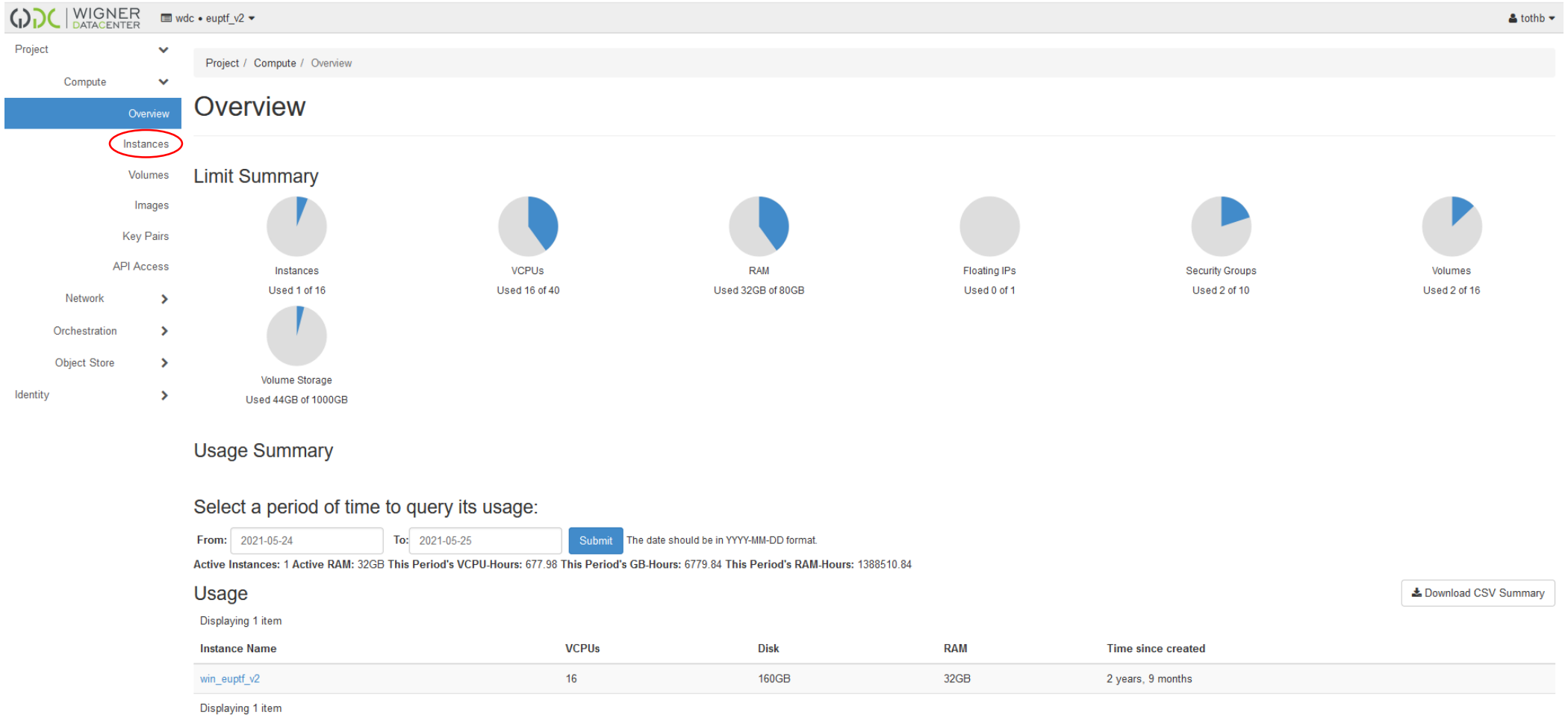
1. lépcső



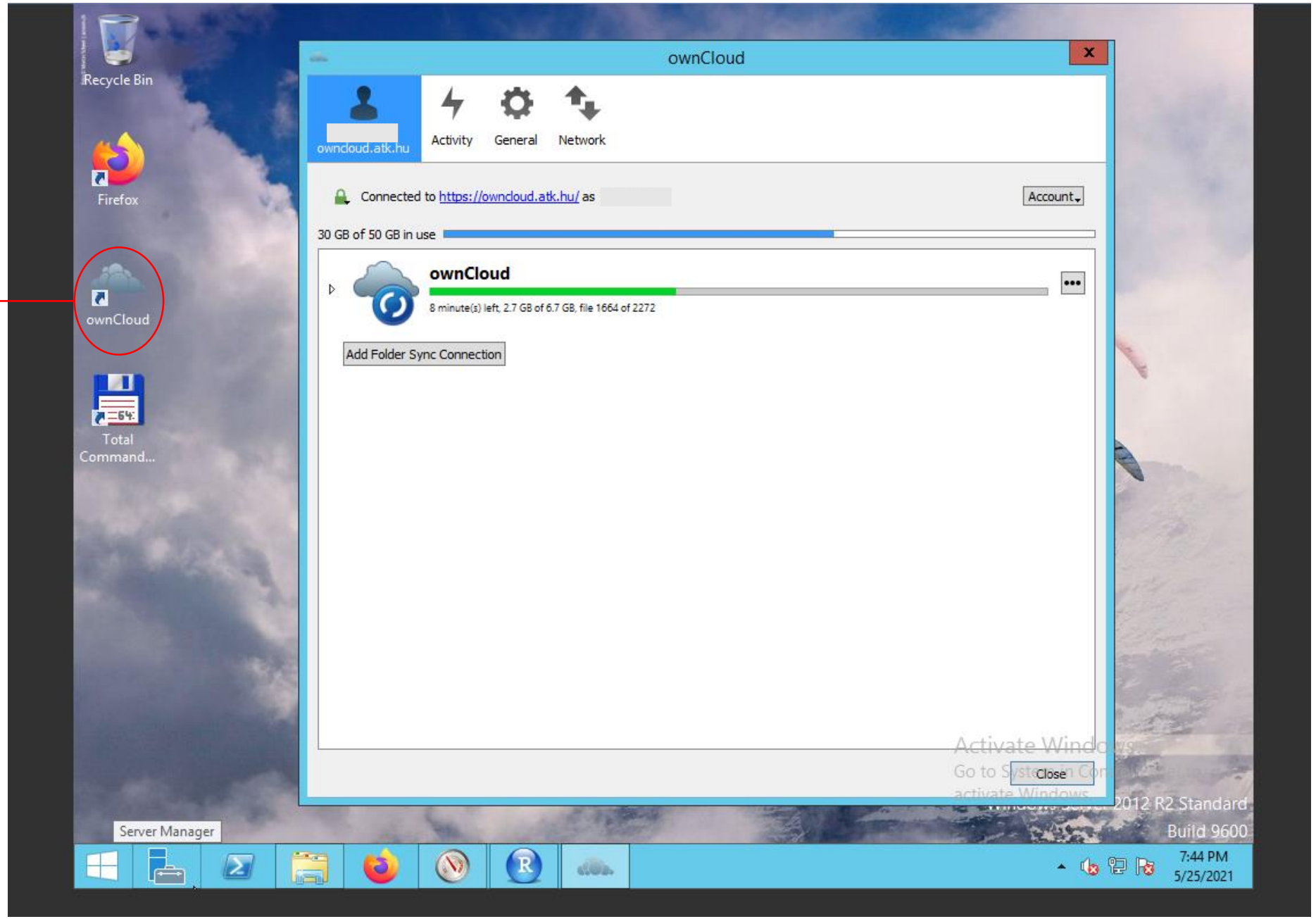
2. lépcső



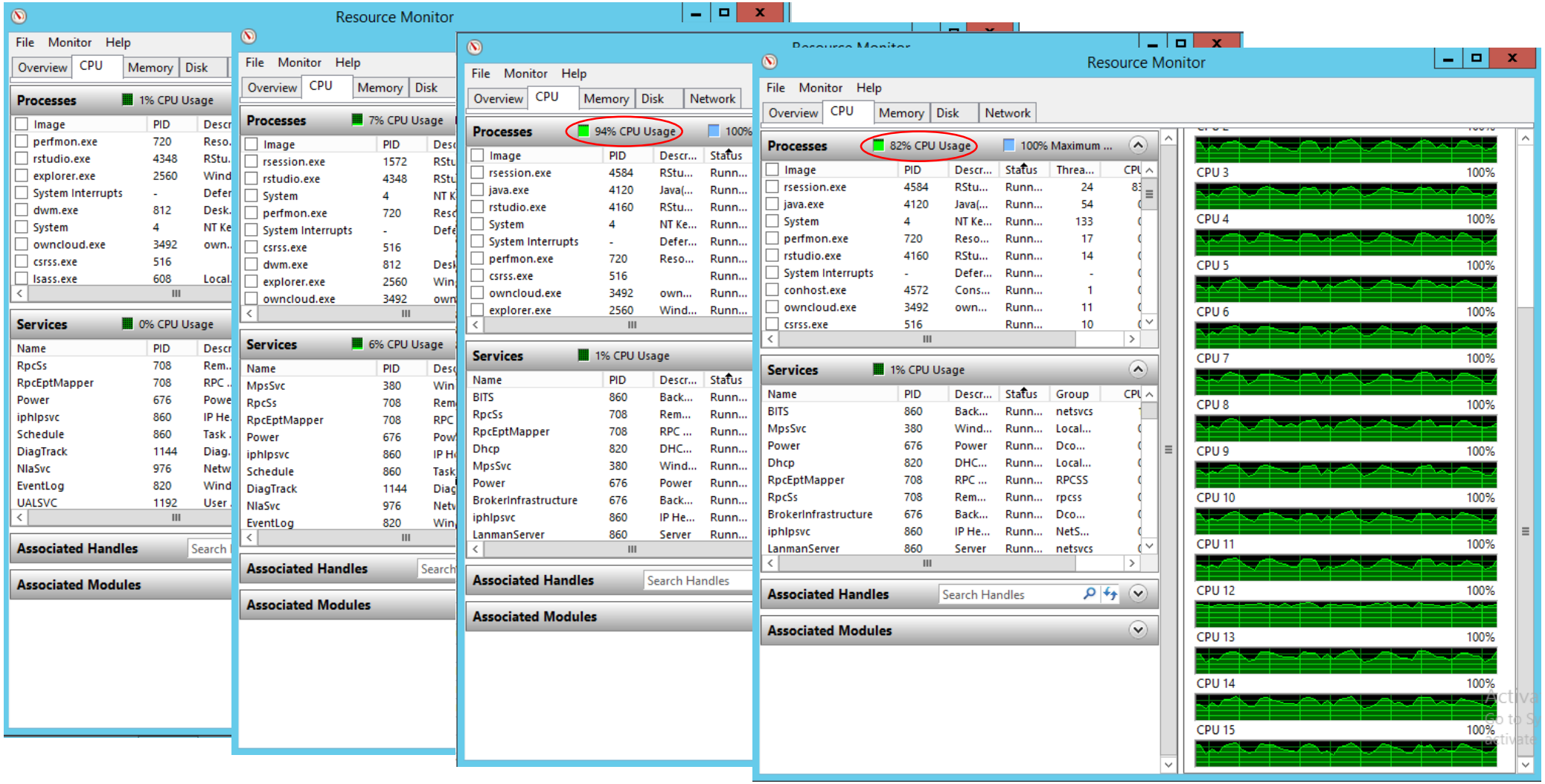
Projekt felülete



Saját gépen látom a létrehozott fájlokat akkor is ha kilépek a WDC-ből.



Erőforrás figyelő



euptfv2

```
1  library(caret)
2  library(parallel)
3  library(ranger)
```

```
> library(parallel)
> detectCores()
[1] 16
```

```
# ----------------------------------------------------------------------
#       euptf ver 2.0
#            Szabo (Toth), B.; Weber, T.K.D.; Weynants, M.
#
#            TUNE the random forest
#
#            author of script: Tobias KD Weber <tobias.weber@posteo.de>
#                              Brigitta Szabo <toth.brigitta@agrar.mta.hu>
#                first version: 20.06.2018.
#                last changes: 03.02.2020.
#  ----------------------------------------------------------------------


source("setupRF.r")
# 1 INITIALISE AND RUN ------------------------------------------------------
################################################################################
#
#
#                INITIALISE  |  GENERAL PROJECT OPTIONS
#
#
################################################################################

for (j in 1:nvar) {

      if(!dir.exists(file.path(getwd(),"results", var.predict[j]))) {dir.create(file.path(getwd(),"results", var.predict[j]))}
      #
      CV.rf.L <- rep( list(vector()), nl )
      ## RUN
      for (i in 1:nl){

            # prepare data sel selector
            if(sum(c("CACO3","PH_H2O","CEC")%in%sel.L[[i]]) == 0){

                  sel.data <- paste("test", var.predict, sep = "")

            } else {
                  sel.data <- paste("test", var.predict, "chem", sep = "")
            }

            # SELECT THE REQUIRED DATASET FOR THE DIFFERENT SUBSETS SPECIFIED IN SEL.V
            tune.dat <- data.ptf[all.rowFun(!is.na(data.ptf[ , c(var.predict[j],sel.L[[i]]) ])) & !data.ptf[ ,sel.data], ]

            ################################################
```

https://github.com/TothSzaboBrigitta/euptfv2/blob/master/point_estimation/THS_FC2_FC_WP_AWC2_AWC/tuneRF.R

```
44            ##################################################
45            #                                                #
46            #           SETUP THE RANDOM FOREST PART         #
47            #                                                #
48            ##################################################
49
50
51            #           TRAIN AND CONTROL
52            tc <- trainControl(method = "repeatedcv",
53                               number = 5,
54                               repeats = 10)
55
56            #           GRIDDING
57            rfGrid <- expand.grid(mtry = seq(2, length(sel.L[[i]]), 1),
58                                  splitrule = c("variance", "extratrees", "maxstat"),
59                                  min.node.size = 10)
60
61            set.seed(1253)
62
63            #           RUNNING THE MODEL
64            CV.rf.L[[i]] <- train(as.formula(paste(var.predict[j], paste(sel.L[[i]], collapse = " + "), sep = " ~ ")),
65                                  data = tune.dat,
66                                  method = "ranger",
67                                  importance = "impurity",
68                                  trControl = tc,
69                                  tuneGrid = rfGrid,
70                                  verbose = TRUE,
71                                  num.trees = 200,
72                                  num.threads = (detectCores()-1))
73
74
75            # remove unused variables
76            rm(tc, rfGrid, sel.data, tune.dat)
77
78        } # end of i loop
```

euptfv2: updated hydraulic pedotransfer functions for Europe

Brigitta Szabó[1], Melanie Weynants[2], Tobias KD Weber[3]

[1]Institute for Soil Sciences, Centre for Agricultural Research, Budapest, Hungary (toth.brigitta@atk.hu), [2]European Commission Joint Research Centre, Ispra, Italy (melanie.weynants@ec.europa.eu), [3]Institute of Soil Science and Land Evaluation, University of Hohenheim, Stuttgart, Germany (tobias.weber@uni-hohenheim.de)

New set of algorithms was derived with random forest method for 32 input combinations, provide built-in prediction uncertainty computation.

Predicted soil hydraulic properties	RMSE based on input combination
Saturated water content (THS)	0.020 – 0.068 $cm^3\ cm^{-3}$
Field capacity at -330 cm matric potential head (FC)	0.046 – 0.055 $cm^3\ cm^{-3}$
at -100 cm matric potential head (FC_2)	0.040 – 0.060 $cm^3\ cm^{-3}$
Wilting point (WP)	0.037 – 0.048 $cm^3\ cm^{-3}$
Plant available water based on FC at -330 cm (AWC)	0.043 – 0.053 $cm^3\ cm^{-3}$
based on FC at -100 cm (AWC_2)	0.045 – 0.060 $cm^3\ cm^{-3}$
Saturated hydraulic conductivity (KS)	0.89 – 1.18 $\log_{10}$(cm day^{-1})
Moisture retention curve (VG)	0.041 – 0.068 $cm^3\ cm^{-3}$
Hydraulic conductivity curve (MVG)	0.61 – 0.71 $\log_{10}$(cm day^{-1})

Availability of the PTFs:

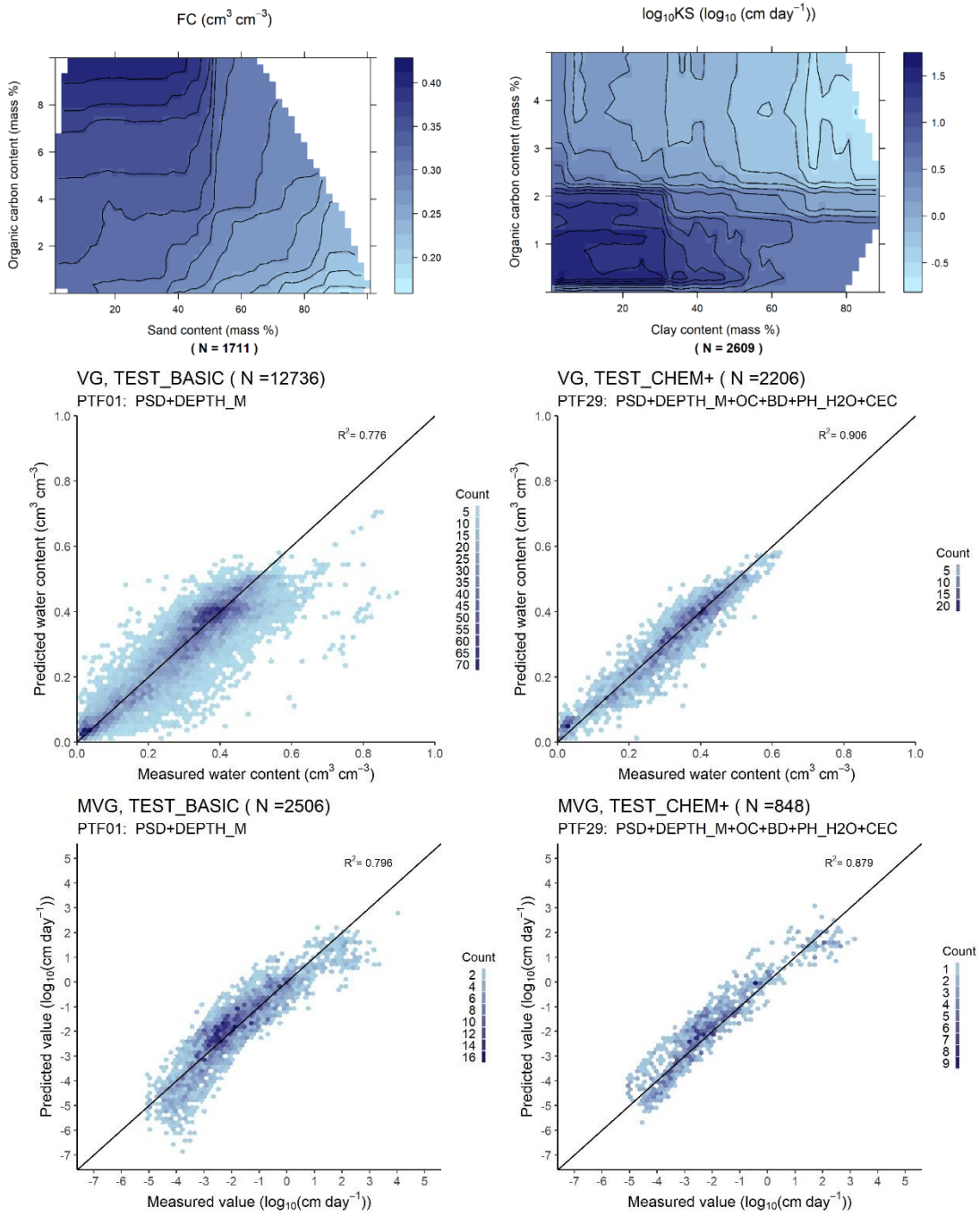
- user friendly web interface: https://ptfinterface.rissac.hu
- R package: https://github.com/tkdweber/euptf2

Szabó, B., Weynants, M., Weber, T.K., 2020. Updated European Hydraulic Pedotransfer Functions with Communicated Uncertainties in the Predicted Variables (euptfv2). Geosci. Model Dev. https://doi.org/10.5194/gmd-2020-36

Szabó, B., Gyurkó, D., Weynants, M., Weber, T.K.D., 2019. Web interface for European hydraulic pedotransfer functions (euptfv2). https://doi.org/10.34977/euptfv2.01

Weber, T. K. D., Weynants, M., and Szabó, B.: R package of updated European hydraulic pedotransfer functions (euptf2), Zenodo, https://doi.org/10.5281/zenodo.4281045

This research has been supported by the Hungarian National Research, Development and Innovation Office (grant no. KH124765), the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (grant no. BO/00088/18/4), and the German Research Foundation (grant no. SFB 1253/1 2017).



On behalf of the János Bolyai Research Scholarship we are grateful for the use of the of ELKH Cloud (https://science-cloud.hu/), which significantly helped us achieve the results published in this paper.

Klaszterezési feladat

library(parallel)
library(h2o)

```
200  localH2O = h2o.init(nthreads = -1)
201  m.hex <- h2o.uploadFile(path = "D:/data_ownCloud/derive_HRU/clustering_RFK_Balaton_cm/input_data
     /rdata_RFK.csv")
202  class(m.hex)
203  summary(m.hex)
204  str(m.hex[, c(2:16)])
205  km2O_est_k <- h2o.kmeans(training_frame = m.hex[, c(2:16)], standardize = TRUE, k = 20, x =
     names(m.hex[, c(2:16)]), nfolds = 5, keep_cross_validation_predictions = TRUE, estimate_k = TRUE
     , max_iterations = 100) # altough the dataset includes NA use = "complete.obs" was deleted from
     the code because that caused an error and clustering stopped
```

```
228  h2o.shutdown()
```

Klaszterezési feladat

library(snowfall)
library(parallel)
library(NbClust)

```
71  i.lst <- c("kl", "ch", "hartigan", "ccc", "scott", "marriot",
    "trcovw", "tracew", "friedman", "rubin", "cindex", "db",
    "silhouette", "duda", "pseudot2", "beale", "ratkowsky",
    "ball", "ptbiserial", "gap", "frey", "mcclain", "gamma",
    "gplus", "tau", "dunn", "hubert", "sdindex", "dindex", "sdbw"
    )
72  # i.lst <- c("kl", "gamma", "gplus", "tau", "dunn", "hubert",
    "sdindex", "dindex", "sdbw")
73
74  sfInit(parallel=TRUE, cpus=detectCores()-1)
75  sfExport("i.lst", "data_z")
76  sfLibrary(NbClust)
77  cl.lst <- sfClusterApplyLB(1:length(i.lst), function(i
    ){NbClust(data_z[sample.int(nrow(data_z), size=5e2),],
    distance="euclidean", min.nc=2, max.nc=20, method="kmeans",
    index=i.lst[i])}) # on 10000 samples it did not finish
    calculations in 10 hours, on 200 samples it goes fast
78  sfStop()
```

Becslések alkalmazása

```
10  library(rgdal)
11  library(euptf)
12  library(snowfall)
13  library(snow) # to run paralell
```

```
136 sfInit(parallel=TRUE, cpus=6)
137 sfLibrary(rgdal)
138 sfLibrary(sp)
139 sfLibrary(raster)
140 sfLibrary(GSIF)
141 sfLibrary(euptf)
142 sfExport("pr.dirs", "predictPTFgrid")
143 out <- sfClusterApplyLB(pr.dirs[17:1370], function(i){try( predictPTFgrid(i, pathIn=".", pathOut
    = "./ptfout") )})
144 sfStop()
```

EU-SoilHydroGrids:

- https://esdac.jrc.ec.europa.eu/content/3d-soil-hydraulic-database-europe-1-km-and-250-m-resolution
- https://www.mta-taki.hu/en/eu_soilhydrogrids_3d

Tóth, B., Weynants, M., Pásztor, L., Hengl, T. 2017. 3D Soil Hydraulic Database of Europe at 250 m resolution. Hydrological Processes. 31:2662–2666.

Becslések alkalmazása

```
 9  require(itertools)
10  # run in parallel on more cores by smaller part of the data
11  require(foreach)
12  # set using more cores:
13  require(doSNOW)
14  require(parallel)
```

```
> detectCores()
[1] 8
> getDoParWorkers()
[1] 1
> cl <- makeCluster(8, type="SOCK")
> registerDoSNOW(cl)
> getDoParWorkers()
[1] 8
```

```
127  # system.time(predictions_1000 <-
128    # foreach(d=isplitRows(rdata[c(1:1000),], chunks=100),
129    #         .combine=c, .packages=c("stats")) %dopar% {
130    #   predict(cforest, newdata=d, type="response")
131    # })
132
133  # system.time(predictions_10000 <-
134    # foreach(d=isplitRows(rdata[c(1:10000),], chunks=1000),
135    #         .combine=c, .packages=c("stats")) %dopar% {
136    #   predict(cforest, newdata=d, type="response")
137    # })
138
139  # system.time(predictions_100000 <-
140    # foreach(d=isplitRows(rdata[c(1:100000),], chunks=1000),
141    #         .combine=c, .packages=c("party")) %dopar% {
142    #   predict(cforest, newdata=d, type="response")
143    # })
144
145  # system.time(predictions_200000 <-
146    # foreach(d=isplitRows(rdata[c(1:200000),], chunks=2000),
147    #         .combine=c, .packages=c("party")) %dopar% {
148    #   predict(cforest, newdata=d, type="response")
149    # })
```

```
190  predictions_100000 <-
191    foreach(d=isplitRows(rdata[c(1:100000),], chunks=1000),
192            .combine=c, .packages=c("party")) %dopar% {
193      predict(cforest, newdata=d, type="response")
194    }
```